Using genome-wide approaches to characterize the relationship between genomic variation and disease:

A case study in Oligodendroglioma and

*Staphylococcus aureus*

by

Nicole V. Johnson

Department of Computational Biology and Bioinformatics
Duke University

Date: ______________________________

Approved:

_______________________________
Lindsay G. Cowell, PhD, Advisor

_______________________________
Vance G. Fowler, MD

_______________________________
Terrence S. Furey, PhD, Chair

_______________________________
Sayan Mukherjee, PhD

_______________________________
William K. Scott, PhD

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology and Bioinformatics in the Graduate School of Duke University

2010
ABSTRACT
(Bioinformatics, 0715)

Using genome-wide approaches to characterize the relationship between genomic variation and disease:
A case study in Oligodendroglioma and
Staphylococcus aureus

by
Nicole V. Johnson

Department of Computational Biology and Bioinformatics
Duke University

Date: __________________________
Approved:

______________________________
Lindsay G. Cowell, PhD, Advisor

______________________________
Vance G. Fowler, MD

______________________________
Terrence S. Furey, PhD, Chair

______________________________
Sayan Mukherjee, PhD

______________________________
William K. Scott, PhD

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology and Bioinformatics in the Graduate School of Duke University
2010
Abstract

Genetic variation is a natural occurrence in the genome that contributes to the phenotypic differences observed between individuals as well as the phenotypic outcomes of various diseases, including infectious disease and cancer. Single nucleotide polymorphisms (SNPs) have been identified as genetic factors influencing host susceptibility to infectious disease while the study of copy number variation (CNV) in various cancers has led to the identification of causal genetic factors influencing tumor formation and severity. In this work, we evaluated the association between genomic variation and disease phenotypes to identify SNPs contributing to host susceptibility in *Staphylococcus aureus* (*S. aureus*) infection and to characterize a nervous system brain tumor, known as oligodendroglioma (OD), using the CNV observed in tumors with varying degree of malignancy.

Using SNP data, we utilized a computational approach, known as in silico haplotype mapping (ISHM), to identify genomic regions significantly associated with susceptibility to *S. aureus* infection in inbred mouse strains. We conducted ISHM on four phenotypes collected from *S. aureus* infected mice and identified genes contained in the significant regions, which were considered to be potential candidate genes. Gene expression studies were then conducted on inbred mice considered to be resistant or susceptible to *S. aureus* infection to identify genes differentially expressed between the two groups, which provided biological validation of the genes identified in significant ISHM regions. Genes identified by both analyses were considered our
top priority genes and known biological information about the genes was used to determine their function roles in susceptibility to *S. aureus* infection.

We then evaluated CNV in subtypes of ODs to characterize the tumors by their genomic aberrations. We conducted array-based comparative genomic hybridization (CGH) on 74 ODs to generate genomic profiles that were classified by tumor grade, providing insight about the genomic changes that typically occur in patients with OD ranging from the less to more severe tumor types. Additionally, smaller genomic intervals with substantial copy number differences between normal and OD DNA samples, known as minimal critical regions (MCRs), were identified among the tumors. The genomic regions with copy number changes were interrogated for genes and assessed for their biological roles in the tumors phenotype and formation. This information was used to assess the validity of using genomic variation in these tumors to further classify these tumors in addition to standard classification techniques.

The studies described in this project demonstrate the utility of using genetic variation to study disease phenotypes and applying computational and experimental techniques to identify the underlying genetic factors contributing to disease pathogenesis. Moreover, the continued development of similar approaches could aid in the development of new diagnostic procedures as well as novel therapeutics for the generation of more personalized treatments.
This is dedicated to the Johnson Family
# Contents

Abstract iv  
List of Tables ix  
List of Figures x  
List of Abbreviations and Symbols xi  
Acknowledgements xii  
Introduction 1  

1 The identification of candidate genes influencing *Staphylococcus aureus* infection 12  

1.1 Project Overview ........................................... 12  
1.2 Methods .................................................. 13  
1.2.1 Survival data ........................................... 13  
1.2.2 Bacterial load in the kidney and peritoneal fluid ........... 14  
1.2.3 Serum levels of interleukin-6 ............................. 15  
1.2.4 In silico haplotype mapping (ISHM) ....................... 15  
1.2.5 ISHM Simulations ...................................... 18  
1.2.6 Gene expression data .................................... 21  
1.2.7 Microarray data analysis ................................. 22  
1.3 Results .................................................... 23  
1.3.1 Phenotype varies with mouse strain for inbred mice infected with *S. aureus* ........................................... 23
1.3.2 ISHM identifies a haplotype block on chromosome 7 significantly associated with kidney bacterial counts . . . . . . . . . 26

1.3.3 Simulations identify an extended region on chromosome 7 containing 40 candidate susceptibility genes . . . . . . . . 28

1.3.4 Eight of the 40 genes in R7 are differentially expressed between susceptible and resistant mice . . . . . . . . . . . . 30

1.3.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32

2 The characterization of genomic aberrations associated with brain tumor severity 39

2.1 Project overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39

2.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

2.2.1 Tumor samples . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

2.2.2 Generation of whole-tiling array platform . . . . . . . . . . . . 42

2.2.3 Sample preparation and hybridization . . . . . . . . . . . . . . 44

2.2.4 Processing and visualization of CGH data . . . . . . . . . . . . 46

2.3 Data analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48

2.4 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50

2.4.1 Summary of genomic variation for all 74 ODs . . . . . . . . . . 50

2.4.2 Unclassified ODs . . . . . . . . . . . . . . . . . . . . . . . . . . . 51

2.4.3 Grade II well-differentiated ODs (WDs) . . . . . . . . . . . . . 52

2.4.4 Grade III anaplastic oligodendroglioma (AOs) . . . . . . . . 53

2.4.5 Evaluation of clone-based genetic variation . . . . . . . . . . . 53

2.4.6 The identification of minimal critical regions (MCRs) . . . . . . 54

2.4.7 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

3 Conclusions 59

Bibliography 67

Biography 83
List of Tables

1.1 Sample sizes for the 13 inbred mouse strains. . . . . . . . . . . . . . . 14
1.2 Permutation analysis identified a genomic region significantly associ-
ated with the kidney count phenotype. . . . . . . . . . . . . . . . . . 27
1.3 Genes identified in block significantly associated with kidney phenotype 31
1.3 Genes identified in block significantly associated with kidney phenotype 32
2.1 The genomic aberrations observed in the 74 oligodendrogliaoma (OD) 
according to tumor grade and chromosome loss of 1p and 19q. . . . . . 54
## List of Figures

1.1 In silico haplotype mapping approach .................................................. 16
1.2 Generation of null and alternative distributions for simulation study. . 19
1.3 Phenotype data from 13 inbred mouse strains injected with *S.*aureus illustrates a wide spectrum of susceptibility to *S.*aureus infection. . . 24
1.4 Median survival obtained from 13 inbred mouse strains injected with *S.*aureus illustrates a wide spectrum of susceptibility to *S.*aureus infection. . . .......................................................... 25
1.5 The power of the method and localization increases with effect size. . 29
2.1 Cellular morphology of oligodendroglioma (OD). . . . . . . . . . . . 41
2.2 The process of conducting comparative genomic hybridization (CGH). 43
2.3 A portion of a hybridized CGH microarray ............................................. 47
2.4 Genome plot of copy number variation. .................................................. 49
2.5 A karyogram of the genomic rearrangements of 74 oligodendroglioma (OD) tumors. .......................................................... 51
List of Abbreviations and Symbols

SNP    Single nucleotide polymorphism
DNA   Deoxyribonucleic acid
DOP-PCR Degenerate oligonucleotide primer polymerase chain reaction
CGH     Comparative genomic hybridization
WHO   World Health Organization
BAC  Bacterial artificial chromosome
PAC P1-derived artificial chromosome
MCR  Minimal critical region
TS    F statistic to identify genomic regions associated with phenotype
OD    Oligodendroglioma
WD    Well-differentiated oligodendroglioma - stage II oligodendroglioma
AO    Anaplastic oligodendroglioma - stage III oligodendroglioma
$maxF_A$ Maximum F-statistic for a single alternative simulation
$maxF_O$ Maximum F-statistic observed across all unique strain patterns
Acknowledgements

I would like to thank my advisor, Dr. Lindsay Cowell, and Dr. Vance Fowler for providing a wonderful training environment in which to conduct this research and their unwavering guidance as mentors. I would also like to thank Dr. Sun hee Ahn, Thomas Rude, and Dr. Hitesh Desmunkh for the experimental support, expertise, and advice provided for the project to study host susceptibility to *Staphylococcus aureus* infection. Additionally, Charlotte Nelson and Dr. William Scott who provided insightful ideas and thoughts on the direction of this project. I would also like to thank Terry Furey and Sayan Mukerjee who have provided guidance throughout the my graduate career and research projects. Last but not least, I would like to thank my family and friends for all of their love and support throughout this process as I would not have made it without them.
Introduction

Susceptibility to disease is influenced by a variety of factors, including genetic factors. Genetic variants have been identified as a main contributor to the development of certain diseases, including infectious diseases and some types of cancer, and to influence the variety of phenotypic outcomes observed. The natural genetic variation in the genome has thus been the basis of disease studies for years and has led to the characterization of various disease phenotypes by their genomic profiles and the identification of causal genetic factors that predispose individuals to disease. The goal of my thesis work is to study the association between genomic variation and disease using two case studies: single nucleotide polymorphisms and their association with susceptibility to Staphylococcus aureus (S. aureus) infection in mice and copy-number variation and its association with tumor progression in the context of oligodendroglioma.

Single nucleotide polymorphisms (SNPs) are the most common type of genomic variants in the genome. SNPs are single base pair mutations that have a frequency of more than 1% in a population of individuals where a single allele is substituted with another. SNPs can reside in non-coding regions where they may influence the gene’s transcription, or the translation and post-translational modification of the gene products. SNPs in promoter or enhancer sequences can influence the level of transcription; SNPs in introns could interfere with mRNA splicing [1]; SNPs in untranslated regions can impact translation, primarily through their effect on mRNA
stability or mRNA processing. SNPs play a prominent role in host genetics of infectious disease, as mutations in immune genes have a direct effect on the success of the immune response (reviewed here [2], [3]) and have also been shown to influence host susceptibility to infection [4][5].

Sequencing efforts have identified millions of SNPs across the human genome [6] that have become the basis of many genome-based approaches to identify genomic variants associated with disease phenotypes. Genome-wide association studies (GWAS) with a case-control design have been conducted in the human to identify genes influencing human disease by detecting SNPs, or other causal variants, associated with diseased individuals in a population. GWAS have gained popularity due to its ability to conduct large-scale genome-wide analyses with increased resolution due to its dense SNP map. GWAS have identified disease risk loci associated with various diseases, including type 2 diabetes, breast cancer, and Alzheimer’s disease [7][8][9]. In recent years, there has been an emerging interest in using GWAS to study infectious disease. Felly et al. [10] conducted GWAS to study infectious disease where they identified two SNP loci influencing the viral load in HIV-positive patients during the asymptomatic phase. Another study used GWAS to investigate susceptibility to infectious disease and identified functionally related genomic variants associated with Kawasaki disease [11]. Although GWAS have proved to be useful in studying infectious disease so far, the platform is limited in its ability to detect rare variants, which would be problematic when multiple rare variants are contributing to susceptibility [12]. Additionally, the current platform design of GWAS tend to be more skewed to common variants, which may may not be ideal for studying infectious disease [12].

Another type of genomic variation commonly observed across the genome are copy number variants (CNVs). Unlike SNPs, the size of CNVs can range from a single nucleotide basepair to thousands of kilobasepairs [13]. CNVs have been shown
to differ between individuals resulting in phenotypic differences [14], but even in the case of identical twins there were differing CNVs[15]. CNVs can arise in the germline during chromosomal crossover events in meiosis, where there is an unbalanced exchange of chromosome material, or somatically, which could be caused by factors, such as environmental carcinogens or non-allelic homologous recombination events in the genome [16]. CNVs consist of deletions (losses) and amplifications (gains) of genomic DNA sequence that result in differences in the number of copies of a DNA sequence, as well as insertions, translocations, and inversions. Chromosomal deletions are genomic alterations in which portions of DNA are lost that may result in the partial or complete loss of genes. Amplifications may result in the duplication of whole genes, or in some cases even more than two copies are created. Copy number changes that overlap with gene regions can result in the under- or over-abundance of gene product [17] and are key contributors to the development of neurological disorders [18][19], susceptibility to various diseases [20][21][22][23], and cancer [24].

Comparative genomic hybridization (CGH) is an experimental technique commonly used to detect copy number variation in the genome [25]. The method is based on the hybridization of two differentially labeled DNA samples whose resulting signal intensities reflect the difference in copy number between a normal reference genome and a test, or diseased, sample. Metaphase-CGH was one of the earlier designs where a fluorescently labeled normal reference and test DNA sample were hybridized to a metaphase chromosome to identify copy number differences, but this method was limited by its poor resolution. Array-based CGH (aCGH) was later developed that utilized a microarray technology that provided increased resolution for detecting chromosomal alterations as compared to the previously used metaphase-CGH. The generation of minimum tile-path arrays provided complete coverage of the human genome using large insert bacterial insert clones (BACs) generated by the Human Genome Project. These BAC-CGH arrays were convenient for the detection of large-
scale CNV, but limited in detecting well-defined CNV regions given the size of the BACs (80-200kb). More recently, oligonucleotide CGH (oligo-CGH) was developed where the arrays used a larger number of hybridization targets, consisting of oligonucleotides 50-60kb in size, that provided increased resolution, accuracy in detecting CNV regions, and greater genome coverage [26]. Despite these improvements, the increase in hybridization targets enhanced the noise-to-signal ratio problem associated with aCGH experiments[27] and is still vulnerable to the effect of poor quality DNA that can affect hybridization [28]. In recent years, there has been increasing interest in using SNP-based association studies to detect CNVs [29] and many of the issues associated with this platform are similar to those of GWAS (e.g. sample size to detect modest alleles) in addition to developing appropriate statistical methods to analyze the data and address specific properties associated with CNVs [30].

Sequencing technologies have also been useful in the identification of CNV. Whole-genome shotgun sequencing was previously used to detect germline copy number variation [31][32] by comparing the number of sequence reads generated for 2 individual genomes where each set of sequences would correspond to the copy number variation present across the genome. Next generation sequencing (NGS) has provided increased accuracy in the detection of well defined CNV regions, which is a limitation to aCGH methods, but sequencing studies are still expensive and this is currently a limiting step in conducting genome-wide analyses across multiple individuals [30].

Large-scale re-sequencing efforts in mice has led to the identification of genetic variation in mice [6] and provided SNP data for numerous inbred mouse strains that is located in the Mouse Genome Informatics (http://www.informatics.jax.org/) and Perlegen SNP (http://mouse.perlegen.com/mouse/index.html) databases. These resources have enabled the development of SNP-based methods for the identification of genetic variants influencing disease phenotypes. GWA studies conducted in the
mouse have used individual SNPs to identify causal variants associated with lung tumor susceptibility [33], schizophrenia [34], and the innate immune response [35]. The availability of SNP data also enabled the development of computational methods, such as in silico haplotype mapping (ISHM). ISHM uses a multi-SNP approach that compares the strains’ SNP allele differences within a pre-defined genomic interval to their corresponding phenotypes to identify phenotype-associated genomic regions [36]. This method has detected genes associated with various phenotypes, such as resistance or susceptibility to malaria [37], hypertension [38], bone mineral density [39], atherosclerosis [40], plasma lipid concentrations in cholesterol [41][42], and albuminuria in age-related kidney disease [43]. Additionally, ISHM has also identified SNPs within large QTLs found through conventional intercross experiments [44][45]. While experimental validation of ISHM-identified genomic regions is necessary, as demonstrated by using gene expression [46] and linkage analysis [47], ISHM offers significant advantages over the use of conventional quantitative trait locus (QTL) mapping alone in the form of reduced time and financial costs and the identification of smaller genomic regions.

Despite the successes of ISHM, the method has been criticized for having a high false positive rates resulting from the recent common ancestry of inbred mouse strains [48] and for having low power [49][50]. Several studies have attempted to improve ISHM methods by accounting for these issues in their models. Kang et al. [48] developed a genome-wide association method, called EMMA, to improve upon the high false positive rates typically associated with ISHM by accounting for the inbred strain relatedness, or population structure, using a mixed-model approach. In addition to this, EMMA also addressed the limited power of ISHM by using phenotype values for individual mice in the calculation of the test for association, instead of the mean or median values for a strain, to increase the sample size and genetic variability. This method, as well as most other studies using ISHM, experimentally assessed
the false positive and false negative rates for particular phenotypes by comparing the significant findings with known QTLs associated with the disease phenotypes \[34][42\] that were identified using QTL mapping. Cladistic approaches \[49\], which rely on phylogenetic reconstruction, have also been developed to account for population structure among the strains. Others have focused on characterizing the power of the method and the dependence of power on i) strain selection, ii) the frequency and effect size of the disease-causing variant, and iii) local patterns of linkage disequilibrium \[50][51\]. While these limitations must be taken into consideration when designing ISHM studies and interpreting results, ISHM has proven useful in identifying novel disease-associated genes, in particular when used in conjunction with experimental techniques.

In this work, I utilized SNPs in the mouse genome and CNV in brain tumors to evaluate the influence of genetic variation on different disease phenotypes. In the first study, I applied a modified version of the Pletcher et al. \[52\] ISHM method to identify genetic factors influencing susceptibility to Staphylococcus aureus (\textit{S.aureus} infection) in inbred mice. In the second, I used aCGH to characterize genomic regions of copy number change in a type of brain cancer, known as oligodendroglioma, with the tumors phenotype. In the following paragraphs, I will describe the biology and the importance of studying these diseases while the methods and results of each study will be described in greater detail in the subsequent chapters.

\textit{S.aureus} is one of the most common bacterial pathogens of the 21st century and has become the 2nd most common cause of bacteremia in both the United States \[53\] and Europe \[54\]. This gram-positive bacterium colonizes the nasal mucosa of most individuals at some point in their lives, but only a small proportion of these individuals develop active infection. In recent years, \textit{S.aureus} infections have become a major concern for several reasons. First, there has been an increase in resistance to the most commonly used antibiotic, methicillin \[55\], and to the last resort antibi-
otic, vancomycin [56], which has resulted in higher mortality rates associated with hospital acquired infections [57][58][59] and have been estimated at 20% [60]. It is for all these reasons that understanding the pathogenesis of *S. aureus* infection is important. Second, these infections, which at one time, were primarily acquired in the hospital setting, have become more commonly community-acquired infections [61]. It is for all these reasons that understanding the pathogenesis of *S. aureus* infection is important.

*S. aureus* infections have been shown to cause a diverse array of clinical conditions ranging from asymptomatic colonization to severe complicated infections, such as bacteremia [62], endocarditis [63][64], hematogenous seeding of deep tissues [65], and patient mortality. Several studies have been conducted to identify risk factors influencing host susceptibility to or the severity of *S. aureus* infections which have led to the identification of 3 contributing factors. Bacterial virulence factors have been shown to influence the severity of *S. aureus* infections [64][66] and patient mortality to *S. aureus* infections [59], but they also do not fully account for all the observed patient outcomes. Clinical studies identified several non-genetic host factors associated with severe *S. aureus* infections, such as the use of intravascular devices [63], hemodialysis catheters [67][68], cardiac surgery [69], and healthcare contact [53]. Evidence of the role of host genetic factors in susceptibility to *S. aureus* infection was shown in studies where higher rates of *S. aureus* infection were observed in isolated populations that are more genetically distinct, due to reduced mating with outside populations, such as the New Zealand Maori [70], Canadian Native Americans [71], and Australian Aboriginals [72]. Familial clusters of recurrent *S. aureus* infections [73][74], and rare genetic conditions characterized by susceptibility to *S. aureus* (e.g., Job Syndrome, Chediak-Higashi Syndrome) [75][76] have also been described. Differential risks of nasal colonization were observed among patients that were *S. aureus* nasal carriers had a three-fold higher risk of acquiring *S. aureus* bacteremia while
non-carriers had higher mortality risk [77]. In addition, several recent studies have demonstrated the importance of host genetics on susceptibility to S. aureus using inbred strains of mice. For example, von Köckritz et al. [78] showed that inbred A/J mice are highly susceptible to S. aureus and that C57BL6/J mice are resistant. Ahn et al. localized this susceptibility to S. aureus in A/J mice to chromosomes 8, 11, and 18 [79]. These studies not only validate the role of host genetics in disease severity, but also demonstrate the utility of using inbred mouse strains to study S. aureus infection.

A major benefit to using inbred mice for the study of host genetic susceptibility to infectious disease is the availability of highly dense, genome-wide polymorphism SNP data. The generation of a mouse model of human disease enables the ability to control the pathogen, genetic variation, and environmental factors [80], and provides a way to measure disease phenotypes not available in the human, such a bacterial colonization of the deep tissues. Different levels of susceptibility have also been demonstrated among inbred mouse strains when infected with various pathogens [81][82][83] in addition to S. aureus [78].

To study the role of host genetics in susceptibility to S. aureus infection, we applied ISHM to four phenotypes measured in inbred mice infected with S. aureus: survival time, bacterial load in the kidney, bacterial load in the peritoneal fluid, and levels of serum IL-6. Using a genome-wide permutation analysis and estimating Type I error from the distribution of the largest test statistics observed, we identified phenotype(s) considered to be significant, or most likely to provide true associations with S. aureus infection, and identified genomic regions significantly associated with the phenotype. The significant regions were interrogated for genes and considered to be potential candidate genes with a role in host susceptibility to S. aureus infection. Next, we conducted a simulation study to evaluate the power and localization of our ISHM method to determine the acuity of this approach in identifying true
causal variants associated with a disease phenotype. Gene expression microarray experiments were then conducted on inbred mouse strains that were susceptible and resistant to *S.aureus* identify genes that are differentially expressed between the two groups of mice 2 hours after infection with *S.aureus*. We used this data to prioritize genes identified in the significant regions identified by ISHM and considered those genes with differential expression patterns a priority for follow-up studies. Therefore, the genes located in the ISHM-identified genomic region that were differentially expressed between susceptible and resistant *S.aureus* infected mice were promising candidates for genes conferring host susceptibility to *S.aureus*.

In the second part of this work, I described the use of CGH to study the patterns of copy number variation, or genomic aberrations, in oligodendroglioma (OD), which is a type of brain cancer. ODs belong to a group of tumors, known as glioma, which consist of malignant tumors commonly located in the central nervous system that differ in tumor location and onset by age, sex, growth potential, and extent of invasiveness [84]. A grading system developed by the World Health Organization classifies glioma from grade I to IV [85], which reflects the tumors malignancy and severity, but ODs are only classified into two tumor subtypes: grade II well-differentiated OD (WDs) and grade III anaplastic ODs (AOs). ODs are initially characterized using histological analysis that visually assesses various characteristics of the tumor, primarily cell morphology, to determine the tumors’ malignancy or grade, but ODs are commonly misclassified due to their heterogeneous nature [86]. Therefore, the identification of genomic alterations in ODs has provided additional support for the further classification of brain tumors [87][88].

Array-CGH is commonly used to identify regions of copy number variation, or genomic aberrations, among ODs and has successfully characterized various genomic alterations with tumor phenotype and the clinical outcome of patients [89] The most common genomic aberration associated with ODs is the simultaneous loss of the short
arm of chromosome 1 (1p) and the loss of the long arm of chromosome 19 (19q), or loss of 1p19q, making it a signature feature of these tumors. The loss of 1p19q was demonstrated to be the result of a whole-arm translocation of chromosome 1 and 19 [90][91] where a derivative chromosome is formed, which consists of the 1p and 19q arms, and is then lost. This results in three chromosomes: 2 complete chromosomes, 1 and 19, and a derivative chromosome, consisting of the long arm of chromosome 1 and the short arm of 19, hence the loss of 1p19q. This event has been suggested to occur due to the strong homology of the centromeric regions of chromosome 1 and 19 [92], but the reason for this event has yet to be fully explained. The loss of 1p19q is considered to be an early event in tumor formation which has been suspected to be influenced by the loss of tumor suppressors located on chromosome 1p [93]. ODs with the loss of 1p19q is associated with better prognosis [91], improved response to radiation and chemotherapy treatment [94][95], and longer survival [96]. The reasons for higher sensitivity to treatment in these tumors is still an open question, but it is suspected that the loss of genes mediating resistance to treatment and epigenetic mechanisms may play a role [91]. When only chromosome 1p or 19q is lost in ODs, it is typically 1p and these tumors are still associated with better overall survival [97][98].

Various other genomic aberrations have been commonly identified among ODs, but there are a few that have been demonstrated to influence the tumor malignancy and patient survival. Studies have demonstrated that the loss of chromosome 10q [89], 9p [97], and 8q [99] have been associated with shorter survival and worse patient prognosis in ODs. The majority of large-scale genomic aberrations observed in ODs are usually deletions, but tumors with amplifications have been associated with poor prognosis [88].

In this study, I utilized aCGH to identify the genomic alterations associated with 74 ODs that were classified into 3 tumor groups: grade II WDs, grade III AOs, and
unclassified ODs. This study was an exploratory analysis to characterize genomic CNVs with tumor grade and the status of the 1p19q loss. Small regions of copy number changes observed across all ODs without the 1p19q loss were identified and interrogated for genes in an attempt to identify pathways that may be involved in the formation and classification of these tumors as ODs given they did not have the characteristic 1p19q loss. This investigation provided a way to confirm what is known about ODs and evidence to support, or refute, the classification of these tumors.

Lastly, I will describe the conclusions obtained for both of the studies conducted, their clinical implications, and ideas for future studies that could build upon the work described here.
The identification of candidate genes influencing *Staphylococcus aureus* infection

1.1 Project Overview

We utilized an in silico haplotype mapping approach (ISHM) to identify genomic regions associated with susceptibility to *S. aureus* infection in mice. We applied a modified version of the Pletcher et al.[52] method of ISHM with the primary modification being to the approach for estimating Type I error. Pletcher et al. [52] conducted a permutation test for each haplotype block and estimated the block-wise Type I error from the distribution of test statistics. In contrast, we conducted genome-wide permutations and estimated genome-wide Type I error from the distribution of the largest test statistics observed genome-wide. The advantages of this approach include that it accounts for the genome-wide nature of the test for association and the lack of independence between blocks (and test statistics).

We used our ISHM approach to test for association between genotype and four phenotypes collected from inbred mice infected with *S. aureus*: survival, bacterial count in the kidney (KC), bacterial count in the peritoneal fluid, and the level of
serum IL6. Across the four phenotypes, we identified a single block on chromosome 7 significantly associated with the KC phenotype. To determine the area on chromosome 7 in which to search for candidate disease genes, we conducted a simulation study to assess the power of the method and the distance between a disease-causing SNP and the closest haplotype block with the maximum F statistic. These results indicated that genes within 380 kilobases (kb) of the block on chromosome 7 should be considered potential candidate genes. In the extended region 762 kb in length, we identified 40 known and predicted genes, 26 of which are members of the extended kallikrein gene family, which are part of the Kallikrein-kinin, or contact pathway. This pathway influences the immune response elicited to bacteria and induces physiological conditions commonly associated with *S.aureus* infection, such as sepsis. The differential gene expression patterns of these genes were used to further evaluate the potential of these genes to influence host susceptibility to *S.aureus* infection.

1.2 Methods

1.2.1 Survival data

Mice from 13 inbred strains (N = 142) (A/J, BALB/cByJ, 129S1/SvImJ, AKR/J, NZW/LacJ, PWD/PhJ, KK/HIJ, FVB/NJ, BTBR T+ tf/J, DBA/2J, C3H/HeJ, NOD/ShiLtJ, and C57BL/6J) were intra-peritoneally injected with $10^7$ CFU/g of the methicillin-susceptible Sanger 476 strain of *S.aureus* (Table 1). Mice were monitored every 8 h for five days, and survival times were recorded in hours. The median survival time was calculated for each strain and used for in silico haplotype association mapping (ISHM). Mice were sacrificed using CO$_2$ asphyxiation if they appeared moribund. Pain/distress was assessed using a numerical scale for the following characteristics: Appearance (0 = normal; 1 = lack of grooming; 2 = rough hair coat; 3 = abnormal posture); Behavior (0 = normal; 1 = moving slowly; 2 = moving slowly,
irregular ambulation; 3 = immobile). A total score (Appearance plus Behavior) of three indicated significant pain and distress and culminated in the early euthanasia of the animal.

1.2.2 Bacterial load in the kidney and peritoneal fluid

A total of 76 mice from 13 strains were injected with *S. aureus* as described above (Table 1.1). Mice were euthanized at 24 hours post infection. Kidneys were collected from all 76 mice; peritoneal fluid was obtained from 68 of the 76 mice. Kidneys and peritoneal fluid were collected and processed as previously described [30]. Briefly, kidneys were collected from euthanized mice and homogenized in phosphate buffered saline (PBS). Kidney homogenate and peritoneal fluid were diluted serially with PBS 10-fold. The serial dilutions (50 µl) were plated on Tryptic Soy Agar plates, incubated (37°C) overnight, and bacterial colonies were counted. For both phenotypes, the average number of colony forming units per either g or ml was calculated for

<table>
<thead>
<tr>
<th>Strain</th>
<th>Survival</th>
<th>Kidney count</th>
<th>Peritoneal fluid</th>
<th>Serum IL6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/J</td>
<td>15</td>
<td>10</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>BALB/CJ</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>129S1SwM/J</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>AKR/J</td>
<td>14</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>NZW/LacJ</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>KKHI/J</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>PWD/PhJ</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>FVB/NJ</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>BTBR t+tfJ</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DBA 2J</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>C3H/HeJ</td>
<td>15</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>NOD/LtJ</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C57BL6/J</td>
<td>10</td>
<td>9</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>142</td>
<td>76</td>
<td>68</td>
<td>46</td>
</tr>
</tbody>
</table>

Table 1.1: This table shows the number of mice used for each strain for each phenotype. The total number of mice used for each phenotype is noted at the bottom of the column.
each strain and used for ISHM.

1.2.3 Serum levels of interleukin-6

A total of 46 mice from 13 strains were injected with S. aureus as described above (Table 1.1). Mice were euthanized at 24 hours post infection. Blood was obtained by intra-cardiac puncture, serum was separated by centrifugation and the amount of total protein was measured using the bicinchoninic acid method kit (Pierce). Serum samples were diluted to ensure equal amounts of total protein, and the levels of interleukin-6 (IL-6) were estimated using an enzyme-linked immunosorbent assay (Due kit, Invitrogen).

1.2.4 In silico haplotype mapping (ISHM)

The procedure used to conduct the in silico haplotype mapping is illustrated in Figure 1.1a. We obtained genotype data were from the Perlegen SNP database (http://mouse.perlegen.com/mouse/download.html) for the 13 inbred mouse strains for which phenotype data were collected. SNPs with genotype information available for all 13 strains (3,260,963 SNPs) were used for ISHM. Overlapping three-SNP windows were formed using three adjacent SNPs, as described in [13]. For each window, mouse strains were assigned to haplotype groups based on the three SNP alleles within the window forming the strain segregation pattern for the window (Figure 1.1b). Adjacent three-SNP windows with the same strain segregation pattern were combined to form a single haplotype block (Figure 1.1c). Haplotype blocks with only a single haplotype were excluded from the analysis. Among the resulting 1,364,341 haplotype blocks, 66,939 unique strain segregation patterns were observed. For each of the four phenotypes, a test of association was conducted for each pattern using
the weighted F-statistic and the largest observed F-statistic was recorded ($max F_o$).

To determine which haplotype blocks demonstrate a significant genotype-phenotype association, a permutation analysis was conducted for each phenotype to generate a null distribution of maximum F statistics ($max F_p$). Each null distribution was generated from 10,000 permutations. For a single permutation, the observed phenotypic values of the individual mice were permuted, the strain averages were calculated, an F statistic was calculated for each of the 66,939 unique strain patterns, and the maximum F-statistic $max F_p$ was captured.

**Figure 1.1:** a) Using SNP data for the 13 inbred mouse strains, 3–SNP windows were formed using 3 adjacent SNP loci. Neighboring 3-SNP windows with the same strain segregation patterns were merged to form haplotype blocks, which were then grouped by their strain patterns resulting in 66,939 unique strain patterns genome-wide. A test for association was then conducted for each unique strain pattern. b) An example of forming 3–SNP windows using 5 mouse strains. Each character string represents a genomic sequence for a single mouse strain, each red column represents a genotyped SNP locus, and the strains are grouped by similar SNP allele patterns for each window, which is called the strain segregation pattern. c) The adjacent 3–SNP windows are merged to form a single haplotype block because they have the same strain segregation patterns.
The weighted F-statistic was computed in the following way:

\[
\text{weighted } F - \text{statistic} = \frac{SS_{bg}}{df_{bg}} \times \frac{SS_{wg}}{df_{wg}}
\]

and

\[
SS_{bg} = \sum w_g n_g (\mu_g - \mu_T)^2
\]

\[
SS_{wg} = \sum w_g n_i (\mu_{gi} - \mu_g)^2
\]

and

\[
df_{BG} = k - 1, df_{WG} = N - k
\]

which consists of the following variables: \(n_g\) is the number of strains with a given haplotype, \(\mu_T\) is the mean of phenotypic values for all strains, \(\mu_g\) is the mean of all phenotypic values for strains with a given haplotype, \(k\) is the number of haplotypes, \(N\) is the total number of strains, and \(w_g\) is the average genetic distance for the strains with a given haplotype. \(w_g\) is computed by averaging the proportion of SNPs with allele differences between each pair of mouse stains. The difference between Pletcher’s [52] F-statistic and ours was in the calculation of \(w_g\). We calculated a global \(w_g\) by dividing the total number of SNP allele differences between each pair of mouse strains by the total number of SNPs genotyped in both strains genome-wide.

To determine which haplotype blocks demonstrate a significant genotype-phenotype association, a permutation analysis was conducted for each phenotype to generate a null distribution of maximum F statistics (max \(F_p\)). Each null distribution was generated from 10,000 permutations. For a single permutation, the observed phenotypic values of the individual mice were permuted, the strain averages were calculated, an F statistic was calculated for each of the 66,939 unique strain patterns, and the maximum F-statistic max \(F_p\) was captured.

For each phenotype, the 95th quantile (max \(F_{PC}\)) of the null max Fp distribution
was used as the significance threshold for identification of significantly associated genomic regions: haplotype blocks for which the observed F-statistic was greater than \( \max F_{PC} \) were deemed to be significantly associated with phenotype. This procedure is illustrated in Figure 1.2a

### 1.2.5 ISHM Simulations

To assess the power and localization of ISHM within the constraints of our study (e.g., sample size and mouse strain panel) and in our genomic region of interest, we conducted a simulation study in which we selected a hypothetical causal SNP, simulated phenotype values for individual mice across the 13 strains, computed F-statistics for the unique strain segregation patterns, assessed the frequency with which an association was detected, and determined the distance between the causal SNP and haplotype blocks significantly associated with phenotype.

We expected power and localization to depend not only on sample size and the specific mouse strain panel used in a ISHM study, but also on the minor allele frequency of the causal SNP and on local genomic properties such as linkage disequilibrium. We therefore selected a hypothetical causal SNP based on criteria corresponding a single significantly associated haplotype block identified by ISHM. Therefore, the minor allele frequency of the selected causal SNP would correspond to the number of strains with minor 3-SNP allele frequency in the block. For example, if the significantly associated haplotype block identified consisted of 3 haplotypes in which 3 strains had the same 3-SNP allele pattern observed for the first haplotype, 8 strains had the same as the second haplotype, and 1 strain had the same allele pattern as the third haplotype, we would select a hypothetical causal SNP with a minor allele frequency of 23% (3/13). Single strain haplotypes were not considered because we do not consider these haplotypes to be sufficiently informative. We then
A) **Null simulations**

For each of 10,000 simulations

- Sample phenotype values $(n = 130)$ from $N(0,1)$
- Calculate the mean phenotype value for each of 13 strains $(n = 10)$
- Calculate an F statistic for each strain segregation pattern $(n = 66,939)$
- Identify max F

Null distribution of max Fs generated from 10,000 simulated max Fs

B) **Alternative simulations**

- Identify a SNP with a given minor allele frequency
- Identify strains with minor allele

For each of 5,000 simulations

- Sample phenotype values
  - Susceptible Strains
    - For strains with minor allele, sample phenotype values from $N(\beta > 0, 1)$
    - $N = 130$ (Total number of phenotype values sampled)
  - Resistant Strains
    - For strains with major allele, sample phenotype values from $N(0, 1)$

- Calculate the mean phenotype value for each of 13 strains $(n = 10)$
- Calculate an F statistic for each strain segregation pattern $(n = 66,939)$
- Identify max F $> 95\%$ quantile of null
- Identify Fs $> 95\%$ quantile from null

**Figure 1.2:** A simulation study was conducted to determine the power of our ISHM method to detect the causal allele and estimate the distance from the blocks deemed significantly associated with the phenotype. 

a) The null distribution was generated by sampling 130 phenotype values from a normal distribution $(N(0,1))$ for each simulation. Average phenotype values were calculated the 13 mouse strains and used in the calculation of an F statistic for each unique strain segregation pattern. For each simulation, an F statistic was calculated for each strain pattern and the maximum F was captured for a total of 10,000 maximum F statistics. The 95% quantile of this distribution was used as the threshold of significance for the alternative simulations.

b) To generate the alternative distributions, a single SNP location that had a minor allele frequency of 23%. Strains with the minor allele for each SNP were identified and their phenotype values were sampled from the alternative distribution, $N(\beta > 0, 1)$, while strains with the major allele had phenotypic values sampled from the normal distribution, $N(0,1)$. A total of 130 phenotype values were sampled for each simulation, the average phenotype values were calculated for each strain, and an F statistic was computed for each unique strain pattern. For each simulation, the maximum F statistic was captured as well as all strain patterns that had an F statistic greater than the 95% quantile determined by the null.
selected a causal SNP within the gene closest to the significantly associated block which was excluded from the haplotype blocks for the simulation study.

We simulated phenotype values under two hypotheses, one in which the selected SNP has no effect on the disease phenotype, and one in which mice with the minor allele of the selected SNP are at increased risk for disease. Under the hypothesis of no difference in disease phenotype between mice with the major and minor allele, we ran 10,000 simulations. For each simulation, 130 phenotype values (10 for each of 13 strains) were sampled from a standard normal distribution $N(0, 1)$, and the average phenotype value was calculated for each strain. These average values were then used to calculate weighted F statistics for each unique strain segregation pattern, as described above. The 95% quantile of the resulting distribution of maximum F statistics ($F_{NC}$) was used as the threshold of significance to identify haplotype blocks with significant association in the alternative simulations described next.

To simulate phenotype values under the hypothesis that mice with the minor allele of the selected SNP are at increased risk for disease, we simulated phenotype values from two normal distributions. For strains with the major allele at the hypothetical causal SNP, we simulated phenotype values from a standard normal $N(0, 1)$ distribution. For strains with the minor allele, we simulated phenotype values from a normal distribution with mean $> 0$, $N(\beta, 1)$, where $\beta$ is the effect size. We conducted 10 sets of alternative simulations using the following effects sizes: $\beta = [.25, .5, .75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3]$. For each effect size, 5,000 simulations were run. For each simulation, the average phenotype value was calculated for each strain, the average values were used to calculate weighted F statistics for each unique strain segregation pattern, as described above, and the maximum F statistic ($\max F_A$) was recorded.

To assess the power of ISHM for each effect size, we calculated the relative frequency of simulations for which an association was detected by dividing the total number of simulations for which $\max F_A > F_{NC}$ by the total number of simulations
To assess localization of ISHM for each effect size, we made the following calculations. We identified alternative simulations for which \( \text{max } F_A > F_{NC} \) and a haplotype block with \( \text{max } F_A \) was located on the same chromosome as the causal SNP. For each of these simulations, two distances from the causal SNP were computed: 1) the distance from the causal SNP to the closest block with \( \text{max } F_A \) and 2) the distance from the causal SNP to any block with \( F_A > F_{NC} \). The average distance across simulations was calculated for each effect size. The procedure of conducting these simulations is illustrated in Figure 1.2

Using the mouse strain panel and sample size in our study, ISHM achieves 80% power in our genomic region of interest at an effect size of \( \beta = 1.5 \). We therefore utilized the average distance between the causal SNP and the closest block with \( F_A > F_{NC} \) at an effect size of 1.5 to determine the boundaries of the genomic region to interrogate for candidate susceptibility genes.

1.2.6 Gene expression data

Gene expression studies were conducted on mice from six of the 13 inbred mouse strains, the three resistant strains (C57BL/6J, NOD/LtJ, and C3H/HeJ) and the three highly susceptible strains that had high average kidney count values (A/J, AKR/J, and BALB/cByJ). Three mice from each strain were infected with \( S. aureus \) as described above, and blood was taken by intra-cardiac puncture two hours after infection. Blood was also taken from three uninfected mice of each strain. Blood was stored at \(-20^\circ\text{C}\). Total RNA was prepared from mouse blood using the Mouse RiboPure Blood RNA isolation kit (Ambion), and globin mRNA was removed using the Globinclear kit (Ambion). All RNA samples passed the quality criteria of the Agilent Bioanalyzer and were used for the analysis. One round of linear amplification
was performed for all samples (Ambion MessageAmp Primier). Biotin-labeled cDNA was hybridized to Affymetrix Mouse Genome 430 2.0 Array-Chips for 16 hours at 45°C following the manufacturers instruction. The arrays were then washed and labeled with streptavidinphycoerythrin (strep-PE), and the signal was amplified using biotinylated antistreptavidin followed by another round of staining with strep-PE. Washing and staining were performed on the Affymetrix fluidics station according to recommended protocols. Labeled gene chips were scanned using an Affymetrix Genechip Scanner 7G.

1.2.7 Microarray data analysis

Preprocessing was conducted using the Robust Multichip Analysis (RMA) [100] implementation in the Bioconductor affy package (http://www.bioconductor.org/), with an additional step to account for differences in probe hybridization resulting from single nucleotide polymorphisms (SNP) between susceptible and resistant mice. The additional step is referred to as SNP masking [101] and is applied after background correction and quantile-quantile normalization but prior to the determination of probeset expression values. Genomic locations hybridized by each probe were obtained from the Ensembl database (http://www.ensembl.org/index.html), and these genomic locations were compared to the locations of SNPs for which at least one susceptible and one resistant strain have different alleles. Probes that hybridize to such SNPs within their target transcripts were excluded from the determination of probeset expression values. SNP masking was applied to 4,258 probesets. Thirty-seven probesets were excluded from further analysis because fewer than four probes were used to estimate the probeset expression value.

We then identified genes in the region(s) significantly associated with phenotype as determined by ISHM. These probesets were analyzed using Analysis of Variance
(ANOVA) to determine whether there were statistically significant differences in the mean expression levels between susceptible and resistant mice. The following generalized linear model was used:

\[ Y_{i,j,k,l} = B_1^i + B_2^j + T_k + S \times T_{k,l} \]  

where \( B_1^i \) corresponds to IVT batch effects, \( B_2^j \) to hybridization batch effects, \( T \) to infection main effects, and \( S \times T \) to strain-infection interaction effects. Two factor levels were used for infection state, uninfected and infected. Two factor levels were used for strain, susceptible and resistant. False discovery rate (FDR) adjusted \( p \) values were calculated using a false discovery rate of 0.1 [102].

1.3 Results

1.3.1 Phenotype varies with mouse strain for inbred mice infected with \textit{S. aureus}

We injected mice from 13 inbred strains (A/J, BALB/cByJ, 129S1/SvImJ, AKR/J, NZW/LacJ, PWD/PhJ, KK/HJ, FVB/NJ, BTBR T+ tf/J, DBA/2J, C3H/HeJ, NOD/ShiLtJ, C57BL/6J) with \textit{S. aureus} and measured values from four different phenotypes: survival time, and, at 24 hours post infection, bacterial load in the kidney, bacterial load in the peritoneal fluid, and serum levels of Interleukin-6 (IL-6). A wide range of values was observed for each of the four phenotypes (Figures 1.3 and 1.4). Survival data were collected by monitoring mice for a period of five days and recording the time to death in hours (Figure 1.4). There were three resistant strains for which none of the mice died (C57BL/6J, NOD/ShiLtJ, and C3H/HeJ), and four highly susceptible strains with median survival times that were 26 hours (A/J, BALB/cByJ, 129S1/SvMJ, and AKR/J). For the remaining six strains, intermediate median survival times were observed.

At 24 hours post infection, the kidney and a peritoneal lavage were taken from
sacrificed mice. Serial dilutions of homogenized kidney and peritoneal lavage were plated, and the number of colony forming units (cfu) counted (Figures 1.3b and 1.3c, respectively). The three resistant strains along with KK/HIJ and FVB/NJ had low average kidney count values (< 10 cfu/g), while three of the highly susceptible strains along with DBA/2J had high average kidney count values (≥244 cfu/g) (Figure 1.3b). The remaining four strains, including the highly susceptible strain 129S1/SvImJ, had average kidney count values ranging from 20 cfu/g to 97 cfu/g (Figure 1.3b). The highest average peritoneal fluid counts were observed for the A/J, 129S1/SvMJ, AKR/J, PWD/PhJ, FVB/NJ, and BTBR T+ tf/J strains (> 1258 cfu/ml), while the BALBcBy/J, NZW/LacJ, KK/HIJ, and C3H/HeJ strains all had very low average peritoneal fluid counts (< 3 cfu/ml) (Figure 1c). The remaining three strains had moderate average peritoneal fluid counts ranging from 33 cfu/ml to 58 cfu/ml (Figure 1.3c).

**Figure 1.3:** The log_{10} of the bacteria colony counts obtained from the kidney (a) and the peritoneal fluid (b), as well as the protein levels of serum IL6 (c), are illustrated for each mouse strain by box and whisker plots. The strains are listed on the x-axis while the bacteria colony counts, for the kidney and peritoneal fluid, and the protein fold change levels, for serum-IL6, are on the y-axis. The extreme values were denoted by the whiskers and the circles represent outliers.
Blood was taken from uninfected mice and from \textit{S. aureus}-infected mice at 24 hours post infection and levels of serum IL-6 were assayed. The fold change level in infected mice relative to uninfected mice of the same strain was computed for each strain, and then averaged (Figure 1.3d). The NOD/ShiLtJ, BALBcBy/J, C57BL/6J, and BTBR T+ tf/J strains had relatively large average fold increases in serum IL-6 levels ($> 8$), while the A/J, 129S1/SvImJ, AKR/J, and DBA/2J strains all had average fold changes that were $< 1$ (Figure 1.3d). The remaining five strains had moderate fold increases ranging from 1 pg/ml to 4 (Figure 1.3d).

When comparing the highly susceptible and resistant strains, we observed that

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{sfig1.png}
\caption{The survival curves for each mouse strain were plotted where the survival times (hours) of the individual mice after infection were on the x-axis and the percentage of mice alive on the y-axis. The strains are listed in the order of susceptibility to \textit{S. aureus} infection where A/J is the most susceptible and C57BL6/J is the most resistant. The extreme values were denoted by the whiskers and the circles represent outliers.}
\end{figure}
the A/J and AKR/J strains have similar phenotype values: low median survival times, high average bacterial counts in the kidney and peritoneal fluid, and low average levels of serum IL-6. Among the three resistant strains, we find that the NOD/ShiLtJ and C57BL/6J strains have similar patterns: no deaths observed within five days post infection, low average kidney counts, moderate average peritoneal fluid counts, and high average levels of serum IL-6.

\subsection*{1.3.2 ISHM identifies a haplotype block on chromosome 7 significantly associated with kidney bacterial counts}

We applied ISHM to each of the four phenotypes to identify associated genomic regions. We inferred haplotype blocks by forming three-SNP windows as described in Pletcher et al. [52] and then forming blocks from adjacent windows with identical patterns of strain three-SNP haplotype assignment. Across the genome, there were 66,900 unique strain-haplotype patterns, each corresponding to 19 haplotype blocks, on average. We calculated a weighted F-statistic for each one, and recorded the maximum observed F statistic max $F_0$.

To determine which blocks were significantly associated with phenotype, we conducted a permutation analysis. For each phenotype, we conducted 10,000 permutations in which we permuted the individual mouse phenotype values, computed a weighted F-statistic for each of the unique strain patterns, and recorded the maximum permutation F-statistic (max $F_p$). Genome-wide type I error was estimated for each $F_o$ by dividing the total number of max $F_p$'s larger than max $F_o$ by the number of permutations (10,000) (Table 1.2). For the kidney count phenotype, max $F_o$ was significant with a p-value (0.0036) much less than the Bonferroni-corrected significance threshold of 0.0125. To identify all blocks in the genome significantly associated with the kidney count phenotype, the 95% quantile of the corresponding max $F_p$
distribution (310.02) was used as a significance threshold. Only a single F statistic exceeded this threshold, the max $F_o = 390.85$, and it corresponded to a single haplotype block 1.9 kb in length located on chromosome 7 at 51,256,409 - 51,258,299 bp (B7). The corresponding strain-haplotype pattern consists of five haplotypes (Hap 1-5) with the following mouse strain assignments:

Hap 1: AKR/J
Hap 2: NZW/LacJ
Hap 3: PWD/PhJ
Hap 4: A/J, BALB/cByJ, DBA/2J
Hap 5: 129S1/SvImJ, BTBR T+ tfJ, C3H/HeJ, C57BL/6J, FVB/NJ, KKHl/J, NOD/ShiLtJ

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>$max F_o$</th>
<th>$max F_p &gt; max F_o$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median survival</td>
<td>91.85</td>
<td>8163</td>
<td>0.8163</td>
</tr>
<tr>
<td>Serum IL6</td>
<td>63.60</td>
<td>4171</td>
<td>0.4171</td>
</tr>
<tr>
<td>Peritoneal fluid</td>
<td>129.29</td>
<td>4442</td>
<td>0.4442</td>
</tr>
<tr>
<td>Kidney</td>
<td>390.86</td>
<td>36</td>
<td>0.0036</td>
</tr>
</tbody>
</table>

Table 1.2: This table shows the results from the ISHM to identify genomic regions associated with phenotype. Column 1 shows the phenotype. Column 2 shows the maximum observed F statistic ($max F_o$). $max F_o$ represents the largest F statistic observed for a unique strain segregation pattern genome-wide using the real phenotypic values for each strain. A permutation analysis was conducted for each phenotype to generate a null distribution of 10,000 maximum F statistics ($max F_p$) using permuted phenotypic values. The number of $max F_p$ that are greater than $max F_o$ is shown in column 3. Column 4 shows the p-values, which were estimated by dividing the total number of $max F_p$ greater than $max F_o$ by the number of permutations (10,000).
Interestingly, nine of the 13 mouse strains, those assigned to haplotypes 2, 3, and 5 above, have an average kidney count value $< 100$ cfu/g. Of the remaining four strains, the A/J, BALB/cByJ, and DBA/2J strains have average kidney counts ranging from 243 cfu/g to 264 cfu/g and share haplotype 4. The AKR/J strain does not share its haplotype with any other strain (haplotype 1) and has an average kidney count value an order of magnitude higher than any other strain.

1.3.3 Simulations identify an extended region on chromosome 7 containing 40 candidate susceptibility genes

To determine the boundaries of the genomic region on chromosome 7 that should be interrogated for candidate disease genes, we conducted simulations to assess the localization of ISHM in the region of $B_7$ and as it depends on the set of mouse strains and sample size used in this study. We expect localization to depend on local patterns of linkage disequilibrium and on the causal SNP minor allele frequency. We therefore selected a hypothetical causal SNP located within the Klk1b11 gene, the gene closest to $B_7$, and with a minor allele frequency of 0.23 corresponding to the minor allele frequency of the $B_7$ haplotype, excluding the haplotypes with only a single strain (haplotype 4). We then ran 5,000 simulations for each of 11 effect sizes, including an effect size of 0 corresponding to the null hypothesis of no effect. For each simulation, we recorded the maximum F-statistic resulting in 11 max F distributions, one null distribution (max $F_{FN}$) and one for each of the 10 non-zero effect sizes (max $F_A$). We identified the 95% quantile for the null distribution (max $F_{FN}$) and used this as a significance threshold.

We evaluated the power of ISHM within our study by determining the number of simulations for which an association was detected ($F_A > F_{NC}$), the number of simulations for which the block containing the causal SNP had an $F_A > F_{NC}$, and the
number of simulations for which the block containing the causal SNP had the max $F_A$ (Figure 1.5). We find that the percentage of simulations for which an association was detected (i.e. any $F_A > F_{NC}$) increased with effect size and exceeded 80% at an effect size of 1.5 (81%). The percentage of simulations that detected an association and for which the block containing the causal SNP has $F_A > F_{NC}$ also increased with effect size and exceeded 80% at an effect size of 1.75 (86%). The percentage of simulations that detected an association and for which the block containing the causal SNP had the max $F_A$ also increased with effect size and reached a plateau near 70% at an effect size of 2.5.

To assess the localization of ISHM at each effect size, we calculated the average

![Figure 1.5](image)

**Figure 1.5**: The power of the ISHM method increased with each effect size and reached 80% power at an effect size of 1.5. The frequency that the causal allele was identified in a significant block also increased with effect size. The frequency of the closest block with the maximum F reached a plateau at 70% while the closest block with a significant F statistic reached 99.9%.

distance between the causal SNP and the closest significant haplotype block or the closest block with the max $F_A$ (Figure 1.5). In both cases, the average distance de-
creased with effect size. For all effect sizes, the average distance between the causal SNP and the closest significant block was shorter than the distance to the closest block with a max \( F_A \). At an effect size of 1.5, the effect size at which ISHM within the constraints of our study achieves 80% power, the average distance from the causal SNP to the closest significant block was 380 kb. We therefore identified a region to interrogate for candidate genes by extending the boundaries of \( B_7 \) by 380 kb in each direction. This resulted in an extended region on chromosome 7,762 kb in length located at 450,876,409 – 51,638,299 bp.

The extended region (\( R_7 \)) contains 54 annotated genomic locations in the Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org/), six DNA segments, eight pseudogenes, and 40 genes (Table 1.3). Of the 40 genes, 26 are members of the kallikrein (Klk) gene family of serine peptidases. The gene within \( R_7 \) that is closest to \( B_7 \) is Klk1b11 located 1.2 kb away. In addition, \( R_7 \) contains the Siglece gene. Siglece and twelve of the 26 Klk genes encode proteins with known functions relevant to the host response to \( S. aureus \) infection (Table 1.3).

1.3.4 Eight of the 40 genes in \( R_7 \) are differentially expressed between susceptible and resistant mice

Gene expression microarray experiments were conducted on mice from six strains, the three resistant strains (C57BL/6J, NOD/LtJ, and C3H/HeJ) and the three highly susceptible strains that also had high average kidney count values (A/J, AKR/J, and BALB/cByJ). Blood was taken from uninfected mice and from mice two hours after infection with \( S. aureus \). Of the 40 genes in \( R_7 \), 38 have probesets on the array. We identified eight that were differentially expressed at the two-hour time point (\( p < 0.05 \) with a false discovery rate of 0.1) (Table 1.3). Klk4 was up-regulated and Klk1, Klk11, Klk1b1, Klk1b11, Klkb8, Klk1b24, and 1700008O03Rik were down-regulated.
Table 1.3: Genes identified in block significantly associated with kidney phenotype

<table>
<thead>
<tr>
<th>Genes</th>
<th>Location</th>
<th>Differential gene expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>KlkB11\textsuperscript{a*}</td>
<td>51251247-51255245</td>
<td>down-regulated</td>
</tr>
<tr>
<td>Klk1\textsuperscript{a*}</td>
<td>51031673-51034633</td>
<td>down-regulated</td>
</tr>
<tr>
<td>KlkB8\textsuperscript{a*}</td>
<td>51206034-51210311</td>
<td>down-regulated</td>
</tr>
<tr>
<td>KlkB1\textsuperscript{a*}</td>
<td>51222138-51226686</td>
<td>down-regulated</td>
</tr>
<tr>
<td>KlkB24\textsuperscript{a*}</td>
<td>51443606-51447822</td>
<td>down-regulated</td>
</tr>
<tr>
<td>Klk1\textsuperscript{a*}</td>
<td>51480820-51484987</td>
<td>down-regulated</td>
</tr>
<tr>
<td>Klk4\textsuperscript{a*}</td>
<td>51136542-51141174</td>
<td>up-regulated</td>
</tr>
<tr>
<td>1700008O03Rik\textsuperscript{a*}</td>
<td>51615413-51641190</td>
<td>down-regulated</td>
</tr>
<tr>
<td>Ceacam18\textsuperscript{a}</td>
<td>50890077-50904280</td>
<td>-</td>
</tr>
<tr>
<td>Siglece\textsuperscript{a}</td>
<td>50906442-50915515</td>
<td>-</td>
</tr>
<tr>
<td>Ctu1\textsuperscript{a}</td>
<td>50927386-50933668</td>
<td>-</td>
</tr>
<tr>
<td>Klk12\textsuperscript{a}</td>
<td>51024267-51028946</td>
<td>-</td>
</tr>
<tr>
<td>Klk10\textsuperscript{a}</td>
<td>51036437-51040779</td>
<td>-</td>
</tr>
<tr>
<td>Klk9\textsuperscript{a}</td>
<td>51047331-51051768</td>
<td>-</td>
</tr>
<tr>
<td>Klk8\textsuperscript{a}</td>
<td>51052947-51059192</td>
<td>-</td>
</tr>
<tr>
<td>Klk7\textsuperscript{a}</td>
<td>51066814-51071729</td>
<td>-</td>
</tr>
<tr>
<td>Klk6\textsuperscript{a}</td>
<td>51079869-51087172</td>
<td>-</td>
</tr>
<tr>
<td>Klk5\textsuperscript{a}</td>
<td>51097639-51106551</td>
<td>-</td>
</tr>
<tr>
<td>KlkB9\textsuperscript{a}</td>
<td>51231431-51235746</td>
<td>-</td>
</tr>
<tr>
<td>KlkB26\textsuperscript{a}</td>
<td>51268048-51272331</td>
<td>-</td>
</tr>
<tr>
<td>KlkB27\textsuperscript{a}</td>
<td>51307660-51312075</td>
<td>-</td>
</tr>
<tr>
<td>KlkB21\textsuperscript{a}</td>
<td>51357662-51361944</td>
<td>-</td>
</tr>
<tr>
<td>KlkB22\textsuperscript{a}</td>
<td>51368043-51372292</td>
<td>-</td>
</tr>
<tr>
<td>KlkB16\textsuperscript{a}</td>
<td>51392137-51396977</td>
<td>-</td>
</tr>
<tr>
<td>KlkB3\textsuperscript{a}</td>
<td>51453561-51457721</td>
<td>-</td>
</tr>
<tr>
<td>KlkB4\textsuperscript{a}</td>
<td>51462805-51467124</td>
<td>-</td>
</tr>
<tr>
<td>KlkB5\textsuperscript{a}</td>
<td>51471844-51476073</td>
<td>-</td>
</tr>
<tr>
<td>1700028J19Rik\textsuperscript{a}</td>
<td>51485303-51491474</td>
<td>-</td>
</tr>
<tr>
<td>2410002F23Rik\textsuperscript{a}</td>
<td>51502092-51518090</td>
<td>-</td>
</tr>
<tr>
<td>Clec11\textsuperscript{a*}</td>
<td>51559642-51562256</td>
<td>-</td>
</tr>
<tr>
<td>Shank1\textsuperscript{a*}</td>
<td>51565623-51613721</td>
<td>-</td>
</tr>
<tr>
<td>2310002F09Rik</td>
<td>50950579-51053315</td>
<td>-</td>
</tr>
<tr>
<td>Acpt</td>
<td>51507518-51512748</td>
<td>-</td>
</tr>
<tr>
<td>Klk14</td>
<td>50945788-50950906</td>
<td>-</td>
</tr>
<tr>
<td>Klk13</td>
<td>50967937-50982128</td>
<td>-</td>
</tr>
<tr>
<td>Klk15</td>
<td>51189141-51194960</td>
<td>-</td>
</tr>
<tr>
<td>Snord88c</td>
<td>51505222-51505311</td>
<td>-</td>
</tr>
<tr>
<td>Snord88a</td>
<td>51505505-51505594</td>
<td>-</td>
</tr>
<tr>
<td>KlkB18-ps</td>
<td>51276190-51276946</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 1.3: Genes identified in block significantly associated with kidney phenotype

<table>
<thead>
<tr>
<th>Genes</th>
<th>Location</th>
<th>Differential gene expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klk1b10-ps</td>
<td>51243161-51247461</td>
<td>-</td>
</tr>
<tr>
<td>Klk1b14-ps</td>
<td>51327230-51331376</td>
<td>-</td>
</tr>
<tr>
<td>Klk1b15-ps</td>
<td>51351788-51354795</td>
<td>-</td>
</tr>
<tr>
<td>Klk1b12-ps</td>
<td>51260009-51264944</td>
<td>-</td>
</tr>
<tr>
<td>Klk2-ps</td>
<td>51170537-51176125</td>
<td>-</td>
</tr>
<tr>
<td>Klk1b7-ps</td>
<td>51200382-51201297</td>
<td>-</td>
</tr>
<tr>
<td>Gm10109</td>
<td>51476387-51482791</td>
<td>-</td>
</tr>
<tr>
<td>Gm15517</td>
<td>51516841-51517261</td>
<td>-</td>
</tr>
<tr>
<td>Gm7238</td>
<td>51555656-51557589</td>
<td>-</td>
</tr>
<tr>
<td>D7Dcr14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D7Mit26.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D7Mit27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D7Mit27.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D7Mit345</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D7Nds5</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Genes with probesets on the microarray

* Significantly differentially expressed

in mice susceptible to *S. aureus* infection relative to those mice that were resistant.

1.3.5 Discussion

Exposure to *S. aureus* has a wide variety of outcomes ranging from asymptomatic colonization to severe and potentially fatal infection. While bacterial [26-28] and environmental factors [29] influence infection severity, there is strong evidence that host genetics also play an important role [9]. Yet the specific host genetic factors contributing to susceptibility are not known. We therefore characterized infection severity in 13 strains of inbred mice and utilized in silico haplotype mapping (ISHM) to identify candidate susceptibility genes. ISHM identified a single significantly associated haplotype block on chromosome 7, and a novel ISHM-based simulation study
identified a larger region containing the block and expected to contain candidate susceptibility genes. This region contains 40 genes, including the 26-member kallikrein gene family.

We characterized infection severity in 13 strains of mice (A/J, BALB/cByJ, 129S1/SvImJ, AKR/J, NZW/LacJ, PWD/PhJ, KK/HJ, FVB/NJ, BTBR T+ tf/J, DBA/2J, C3H/HeJ, NOD/ShiLtJ, and C57BL/6J) by quantifying four disease phenotypes (Figure 1.3). Mice were infected with the Sanger 476 strain of S. aureus and survival time was determined for up to five days. In addition, the extent of bacterial colonization in the kidneys and peritoneal fluid and serum levels of interleukin-6 (IL-6) were assayed at 24 hours post infection. We observed a wide variety of phenotype values across the 13 strains of mice (Figure 1.1). Interestingly, we observe that two of the highly susceptible mouse strains with median survival times of < 26 hours (A/J and AKR/J) also had high average bacterial counts in the kidney and peritoneal fluid, and low average levels of serum IL-6 24 hours after infection. Conversely, two of the highly resistant mouse strains (NOD/ShiLtJ and C57BL/6J) with no deaths within the five-day monitoring period, had low average kidney counts, moderate average peritoneal fluid counts, and high average levels of serum IL-6. High bacterial counts in the kidneys are indicative of bacterial dissemination and are correlated with kidney dysfunction in mouse strains susceptible to S. aureus infection [78]. High bacterial counts in the peritoneal fluid are indicative of deficient host clearing of bacteria, and indeed, von Köckritz et al. [78] recently demonstrated that i) inhibition of neutrophil recruitment to the site of infection renders mice susceptible to infection, and that ii) susceptible and resistant mice show different kinetic profiles in their expression of the neutrophil chemoattractants CXCL1 and CXCL2. Similarly, low levels of IL-6 may be indicative of a deficient host immune response, as IL-6 is an important mediator of inflammation and activator of neutrophils, and has been shown to be required for successful defense against bacterial pathogens, such as Streptococcus pneumoniae.
and Listeria monocytogenes. Taken together, these data indicate that the highly susceptible strains do not effectively clear *S. aureus* from the site of infection, do not mount an effective IL-6-mediated inflammatory response, and are subject to extensive bacterial dissemination. The mechanisms underlying these phenotypes are not yet clear.

We applied ISHM as described in [13] to the four phenotypes and identified one significant association which was between bacterial colonization of the kidney and one haplotype block located at positions 51,256,409–51,258,299 bp on chromosome 7 (*B*7). While ISHM has been criticized for spurious associations [12, 32], there is strong experimental evidence that *B*7 is in fact linked to one or more causal variants: in a separate study conducted by our group the presence of a causal variant on chromosome 7 was demonstrated using consomic mice created from the highly susceptible A/J strain and the highly resistant C57BL/6J strain. Mice from a consomic mouse strain created by replacing the C57BL/6J chromosome 7 with the A/J chromosome 7 [33] were susceptible to *S. aureus* infection with a median survival time of 2.5 days [79].

The Ahn et al. study [79] also showed that A/J chromosomes 8, 11, and 18 confer susceptibility to *S. aureus* when substituted into C57BL/6J mice, yet we do not detect any significantly associated haplotype blocks on these chromosomes. We also do not detect significant associations for the survival, peritoneal fluid, and serum IL-6 phenotypes. One possible explanation for this result is that the current study lacks sufficient power to detect the corresponding causal variants. The mouse strain panel used in this study is relatively small. While some studies have used similar numbers of strains and detected significant associations using ISHM [35], many other studies have used much larger numbers of strains. It is thus likely that with a larger mouse strain panel, additional associations would be detected. A second possible explanation is that the phenotypic similarities observed between mice of different strains
do not result from an underlying genetic variant common to the different strains. The phenotypes we are measuring, in particular the survival and serum IL-6 phenotypes, are quite complex, and it is likely that the phenotypes result from complex interactions between many different genetic variants, each of which is only shared by some subset of the strains exhibiting similar phenotypes. Such genetic variants are unlikely to be detected by ISHM. Finally, the survival times are right-censored, and thus the median survival times of the highly resistant strains are underestimated. This results in an underestimate in the F-statistic sum of squares for the within- and between-haplotype group variability. While underestimates of the within-group variability may result in false positive associations, underestimates of the between group variability may result in false negative associations.

One problematic feature of genotype-phenotype association studies is that the variant(s) contributing to the disease phenotype, as well as the gene(s) affected by it, may be located some distance away from the SNP or haplotype that is detected by the test for association [12, 22, 25]. We therefore developed a novel simulation approach to address this issue by estimating the localization of any significant SNP or haplotype to a causal variant and using this estimate of localization to identify a larger genomic region containing the significantly associated SNP or haplotype and that should be interrogated for candidate disease genes. Using our simulation approach, we identified a 762-kb genomic region (R7) located at positions 50,876,409–51,638,299 bp on chromosome 7. R7 contains 40 known and predicted genes, including all 26 members of the extended kallikrein gene family. Thus, while the use of ISHM alone did not in our case identify a set of candidate genes, the incorporation of the simulations into the ISHM approach allowed us to identify a promising list of candidate genes for continued study.

The utility of the simulation approach goes beyond the current study. The simulation approach estimates the localization of ISHM using features of a particular
study: the mouse strain panel and the number of individual mice phenotyped for each strain. In addition to depending on these features of study design, localization will depend on genomic properties in a particular region of interest (i.e. area of the significant association and/or of the causal variant), such as linkage disequilibrium, and on the minor allele frequency of the causal variant. Because these cannot be known prospectively, our approach is to use features of the ISHM result to guide selection of a hypothetical causal SNP, based on its location and allele frequencies, for use in the simulations. The simulation approach is therefore tailored to a specific study in order to address the question: what is the expected distance between a significant ISHM result and the causal variant(s) driving the association. The simulation approach can thus be incorporated into any future ISHM study and will be particularly useful for investigators applying ISHM to phenotypes for which there are no conventional QTLs that can be used to guide the ISHM search. We recommend incorporation of this simulation approach into future studies where ISHM is applied genome wide.

$R_7$ contained all 26 members of the extended kallikrein gene family, and these genes are promising candidate genes for susceptibility to $S$. aureus infection. Products of the kallikrein genes (KLKs) have well-established roles in the degradation of extracellular matrix (ECM), the generation of antimicrobial peptides, and the regulation of immune responses, particularly inflammation (reviewed in [105],[36,37]). Degradation of ECM facilitates the infiltration by immune cells of the skin and other tissues. Thus, enhanced ECM degradation by KLK enzymes could facilitate $S$. aureus dissemination. Conversely, reduced ECM degradation by KLK enzymes could inhibit the host immune response. Furthermore, the proteolytic activities of KLKs are important for the generation of antimicrobial peptides, particularly cathelicidins and defensins, that directly kill microbes as well as influence innate immune response processes. In this case also, reduced KLK activity would weaken the host immune
response to \textit{S. aureus}.

KLKs play an important role in the regulation of inflammation, particularly through activation of the IL-1 precursor and the potent vasoactive peptides bradykinin and kallidin (refs). Thus, alterations of KLK activity could result in dysregulation of the inflammatory response. We and von K"ockritz et al. [78] observed increased kidney infiltration by \textit{S. aureus} in susceptible mice relative to resistant mice. In addition, von K"ockritz et al. [78] observed increased lung infiltration by \textit{S. aureus} and erythrocytes, as well as evidence of extensive lung hemorrhage in A/J mice. These observations are all consistent with severe, increased microvascular permeability in susceptible mice in response to \textit{S. aureus} infection. von K"ockritz et al. [78] further observed increased levels of serum bradykinin in A/J mice as well as decreased activated partial thromboplastin time. Both observations provide evidence that susceptible mice experience increased microvascular permeability as a result of increased activation of the kallikrein-kinin or contact system. Given the role of KLKs, in particular KLK1, in the production of bradykinin and kallidin, these observations support a hypothesis in which susceptible mice are genetically predisposed to increased KLK activity which results in increased production of bradykinin and kallidin, which in turn results in severe increased microvascular permeability, bacterial dissemination, and sepsis.

To determine whether any of the 40 genes in $R_7$ are differentially expressed between susceptible and resistant mice, we conducted gene expression microarray experiments on mice from six strains: the C57BL/6J, NOD/LtJ, and C3H/HeJ strains which have median survival $>120$ hours and low average kidney bacterial counts 24 hours after infection, and the A/J, AKR/J, and BALB/cByJ strains which have median survival $<26$ hours and high average bacterial counts in the kidney 24 hours after infection. Of the 40 genes in $R_7$, 38 have probesets on the Affymetrix Mouse 430 2.0 array, and we determined that two hours after infection with \textit{S. aureus}, eight
of the 38 genes are differentially expressed between the highly susceptible and highly resistant mice. Seven of the eight genes are members of the extended kallikrein family (Klk1, Klk1b1, Klk1b8, Klk1b11, Klk1b24, Klk4, and Klk11), providing further experimental evidence for the role of these genes in mediating host susceptibility to *S. aureus* infection. The eighth differentially expressed gene is the predicted gene 1700008O03Rik, which has no known function. Klk4 is up-regulated in susceptible mice, and the remaining differentially expressed Klk genes are down-regulated. These results support a hypothesis in which KLK activity is reduced in susceptible mice resulting in a deficient host immune response through decreased degradation of ECM and decreased activation of antimicrobial peptides, as discussed above. Given the large number of Klk genes, however, the varied and wide-ranging functions of their gene products, and the complicated patterns of co-regulation via reciprocal- and auto-proteolysis, a series of gene-specific experiments at the nucleic acid and protein level are required for each of the 26 genes in order to disentangle their precise role in host susceptibility to *S. aureus*. 
The characterization of genomic aberrations associated with brain tumor severity

2.1 Project overview

Oligodendroglioma (OD) are nervous system brain tumors that are routinely characterized by tumor morphology and genomic aberrations. The observed patterns of aberrations can be associated with the tumor’s grade and malignancy, as defined by histological analysis, for the further classification of the tumor phenotype. ODs are typically classified into two tumor subtypes: grade II well-differentiated OD (WDs) and grade III anaplastic ODs (AOs), as defined below. For this project, array-based comparative genomic hybridization (CGH) was conducted on 74 oligodendroglioma (OD) to identify patterns of genomic variation, specifically deletions (losses) and amplifications (gains) of genomic sequence, associated with this tumor type and its two subtypes. Additionally, CGH was conducted on 27 normal, or cancer free, DNA samples to determine the normal copy variation in disease free individuals, which was then used to identify smaller genomic intervals with substantial copy number differences, known as minimal critical regions (MCRs), when compared to the OD
samples. Common genomic rearrangements were identified within and across the sets of tumor samples to generate genomic profiles that were classified by tumor grade and stage of progression. For those samples assigned a grade, a comparison of the tumor aberrations with the tumor grade provided insight about the genomic changes that typically occur in patients with OD ranging from the less to more severe tumor types. The identification of previously identified genomic aberrations associated with ODs, such as the characteristic loss of chromosome 1p and 19q, will provide additional confirmation that the tumor samples were classified as the correct tumor type. The data obtained from the CGH experiments was processed using the Bluefuse software that provided gene annotation for each of the clones contained on the array. Using this information, I identified genes in the altered genomic regions, along with their biological function using online databases, to evaluate the potential consequences of gene duplication or loss. This information was used to generate a list of candidate disease genes for investigation in future experimental studies.

2.2 Methods

2.2.1 Tumor samples

Brain tumor samples were obtained by Dr. Darrell Bigner, Department of Pathology, of the Duke University Medical Center, from patients clinically determined to have OD. Histopathological analysis was conducted on the samples to determine the tumor grades, as defined by World Health Organization (WHO). Tumor samples are stained and visually inspected to assess the cells' morphology [84]. Grade II OD tumor cells have rounded nuclei amidst blood vessels described as having a “chicken wire” appearance (Figure 2.1a), while grade III OD are described as having nuclei that are atypical in shape, or multinucleated, accompanied by increased cell density and necrosis [106] (Figure 2.1b). The grading system is a scale that corresponds to
tumor malignancy and consists of 4 grades: Grade I tumors have low proliferative potential and can possibly be cured by surgery alone, Grade II tumors are likely to progress to higher grades of malignancy, Grade III tumors show evidence of malignancy and must be treated with radiation and/or chemotherapy, and Grade IV tumors are malignant, highly incurable due to pre- and post-operative occurrences and are mostly fatal [85]. This grading system is used across a variety of tumors, but only grades II and III are used for oligodendroglioma. A subset of the samples analyzed here could not be classified as grade II or III, and are therefore analyzed as a separate group. There were a total of 74 ODs that were histologically defined; 30 were classified as grade II well-differentiated (WD) and 30 were classified as stage III anaplastic oligodendroglioma (AO). There were 14 samples that could not be further

Figure 2.1: This image illustrates the difference in cellular morphology of a grade II and a grade III. a) The cells of grade II, or low-grade OD, have low-density and are uniformly round to oval shaped with a round nuclei and are surrounded by a network of blood vessels, commonly called the “chicken wire pattern”, due to its appearance. b) Cells of grade III, or anaplastic OD, are highly dense, have atypical nuclei, and are associated with tissue necrosis. The arrows in the picture are pointing to tissue necrosis, which consists of dead cells and tissue.
classified. Tumor DNA was extracted from tissue using the PureGene DNA isolation kit.

2.2.2 Generation of whole-tiling array platform

The high-resolution comparative genomic hybridization (CGH) arrays used for the CGH experiments were generated in Dr. Simon Gregory’s laboratory at the Center for Human Genetics at Duke University. The CGH arrays were designed to provide genomewide coverage by utilizing complete sets of overlapping, minimum-tile path clones generated by the Human Genome Project, as illustrated in Figure 2.2a. The clones on the array consisted of two types of large insert clones, Bacterial Artificial Chromosome (BAC) and P1-derived Artificial Chromosomes (PAC) clones, which were obtained from the Sanger Institute. The clones were created by inserting a portion of human genomic DNA into a plasmid, which is a stretch of DNA that naturally occurs in bacteria, and these plasmids are then inserted into bacterial cells using a technique referred to as transfection. The plasmids carry genes that are resistant to particular antibiotics and are used as a marker to determine if a plasmid has been inserted. The bacteria with the plasmids are then exposed to an antibiotic, as described below, and only those bacterial cells with the plasmid will survive because they will be resistant to the antibiotic. Bacterial DNA is then used for the creation of the array. I was responsible for completing this process for 4 chromosomes on the array.

The bacterial clone DNA was prepared in our laboratory using the procedure described in [107]. Briefly, bacterial clones were transferred to 96-well deep well plates into bacterial growth media (2X TY). The appropriate antibiotics (kanamycin for PACs and chloramphenicol for BACs) were added to the clone cultures, and the bacteria were grown for 18 hours at 37°C with gentle shaking. Next, the bacterial
cultures were transferred to clean microtitre plates (Greiner) and spun down by centrifugation (Sorvall RT7, Du Pont Company Sorvall, Delaware US) resulting in a pellet of cells. The excess supernatant was discarded, the pellet was collected, and re-suspended in a buffer solution. The clone DNA was precipitated and filtered from the solution using filter plates.

Degenerate oligonucleotide primer polymerase chain reaction (DOP-PCR) was then used to amplify the clone DNA. This produced substantially larger quantities of DNA that was later spotted onto the array to enhance the signal intensities of the

**Figure 2.2:** a) BAC clones are generated using the genomic sequence of human DNA. The minimum tiling path created by these clones provides coverage across the entire human genome where each spot on the array corresponds to a particular clone. B) Normal, or tumor free, human DNA and tumor DNA are differentially labeled, combined, and hybridized to the microarray platform. C) The resulting color of the fluorescent intensities of the clones represent the type of chromosome aberration at the location where green represents duplicated copy number variation (gain), red is a deletion of the genomic sequence (loss), and yellow is normal copy number variation of the tumor sample when compared to the normal sample. The BAC clones were annotated with their corresponding genes and could be correlated with the copy number variation observed at the genomic locations.
tumor DNA hybridized to the clone DNA [108]. The DOP-PCR procedure used for this project is described in Fiegler et al. [6]. In brief, three primers were designed for a targeted amplification of the human DNA sequence with the following sequences:

DOP 1 (CCGACTCGAGNNNNNATGTGG)
DOP 2 (CCGACTCGAGNNNNNCTAGAA)
DOP 3(CCGACTCGAGNNNNNTTCTAG)

The primers were designed to efficiently anneal to human DNA and not E. coli DNA, which is a common contaminant of DNA preparations [6]. The degenerate nature of the primer sequences allows them to bind to multiple locations in the clone DNA sequence, which facilitates the generation of DNA products that will provide adequate representation of human DNA sequence.

After the DOP-PCR, another round of PCR was conducted with a set of primers, called amino-linkers. The amino linkers were designed with a 3’ end sequence that was complementary to the 10-nucleotide bases on the 5’ end of the DOP-PCR primer sequence and were annealed to the clone DNA at this location. The biochemical properties of the amino linker enabled it to have an affinity for attaching to the microarray slides via an amino-binding solution to attach the BAC and PAC DNA to the microarray slide. Lastly, other members in the laboratory printed the DOP-amino linked PCR products onto glass microarray slides (Genetix QArray2 Arrayer) where the DNA products in each spot corresponded to the DNA sequence of a single BAC clone.

2.2.3 Sample preparation and hybridization

For a single CGH experiment, two DNA samples, a normal (disease free) and tumor DNA sample, were hybridized to a single array to identify genomic aberrations
present in the tumor. CGH was conducted on each of the normal control samples to ensure that the individual was genetically stable and there were no chromosomal gains or losses. The two DNA samples hybridized for each CGH experiment were a sex mis-matched pair where one DNA sample was from a female and the other DNA sample was from a male. For example, in the 74 tumor versus control CGH experiments, a tumor sample was combined with a normal sample where each sample was taken from a patient of the opposite sex e.g. a tumor sample from a female patient was mixed with a male control sample. The two DNA samples were differentially labeled with a fluorescent dye where one sample was labeled with Cy3-CTP (Red) and the other with Cy5-CTP (green) (Figure 2.2b). For these experiments, the tumor DNA was always labeled with red and the normal DNA, or control, with green. Given that there was limited normal DNA for each control sample, the same control was not used for all of CGH experiments, but each control used was from a collection of normal DNA samples used to generate a set of control CGH experiments. There were 27 control CGH experiments conducted where a normal, or tumor free, male and female sample were differentially labeled with the fluorescent dyes and hybridized to the array.

Each male-female combined sample was then hybridized to a single microarray slide using the MAUI hybridization station (BioMicro Systems Inc). By hybridizing a pair of samples from patients of the opposite sex together, the X and Y chromosome served as a control for the experiment and confirmation of the sex of the samples based on the sex chromosomes. Since the female has two X chromosomes, there should be a gain of the X chromosome and a loss of the Y chromosome, which reflects the extra copy of the X chromosome and the absence of the Y chromosome in the female sample. After hybridization, the microarray slides were washed to remove the excess hybridization buffer by soaking the slides in a warmed formamide solution, then rinsing in water with chemical detergents, and drying in the centrifuge. During
this post-processing step, the slides were kept away from the sunlight to protect the light-sensitive fluorescent dyes.

2.2.4 Processing and visualization of CGH data

After hybridization of the sample to the slides, the microarray slides were scanned using the Genepix 4100A scanner (Molecular Dynamics) and the fluorescent intensities of the spots were estimated. The overall color of each spot on the array represented the relative signal intensities of the two differentially labeled DNA samples where red represents the loss or a portion of a chromosome (deletion), green being the gain of a portion of the chromosome (duplication), while yellow represents neither a gain or loss of a chromosome portion (Figure 2.3). The fluorescent intensities of the spots on the array also indicate how well the samples hybridized to the microarrays. If there was poor hybridization, the spots were usually very small or the spots’ fluorescent intensities were very low because the tumor-normal sample did not bind to the clone DNA fragments on the microarray. The files obtained from this analysis consisted of the raw intensity values for each of the BAC clones on the array.

Next, I used the BlueFuse microarray analysis software (http://www.cambridgebluegnome.com/) to process the raw data files obtained from the GenePix Scanner. The software determined the quality of the spots using a Bayesian approach by calculating a probability for each spot. Using Bayes theorem, a prior was estimated based on attributes collected from the scanned microarray, such as dye chemistry and probe distribution, and other factors that are less predictable, such as the effects of washing the microarray, were predicted using information obtained from the microarray data. The probability calculated for each clone represented the most likely quality of the spot, given the current data and what is known about the other experimental factors that could affect the intensity, where 0 meant
Figure 2.3: Each dot on the image represents a single spot on a CGH microarray for a single genomic sequence. The majority of the spots are yellow, which suggests this is a genomic region with relatively normal copy number variation. The presence of several small spots indicates that the hybridization efficiency for this particular experiment was fair. The red spot on the array indicates that there was a chromosomal deletion (loss) while the green spot indicates a duplicated region (gain).
the spot was considered to be acceptable and 1 was not. Confidence flag estimates are associated with the 0 - 1 interval in increments of 0.18, where A corresponded to 0 - 0.18, B corresponded to 0.18 - 0.26, and so on, up to 1. I used the quality scores of the clones to filter the results of my CGH experiments, which is described later in the text.

For each spot, the $\log_2$ ratio of the differentially labeled dye intensities were used to determine which of the genomic changes were amplifications or deletions across the genome. Using the ratios for each clone, whole genome plots were generated for each microarray experiment to visualize the large-scale genomic aberrations occurring in each tumor sample, as shown in Figure 2.4. By plotting the log ratio of each clone against the clone’s location, you are able to see chromosomal duplications, regions where the clone intensity ratio is greater than 0, and those with losses, where the ratio is less than 0. These plots were used to identify large genomic aberrations, usually on the order of hundreds of adjacent clones, in each of the tumor samples. The set of 27 control CGH experiments conducted were also analyzed using the Bluefuse software. This set of experiments was used to determine the copy number variation of normal, tumor free, individuals and provided a baseline for the comparison of each of the clone’s log ratios with those of the tumor samples.

2.3 Data analysis

To evaluate the copy number variation in the tumors on a finer scale, I analyzed the $\log_2$ ratio of the intensities for the individual clones to identify small genomic changes occurring in all of the ODs. The data files with all of the clone intensity values on the array and log ratios were obtained from the Bluefuse software. The quality score assigned to each of the clones was provided in the files and those clones with E's and F's for more than 50% of the tumor sample experiments were removed.
from the analysis. After excluding these clones, there were a total of 24,170 clones analyzed. There were no clones for which 50% of the control samples experiments received an E or an F. Using the processed set of control CGH experiments, the median intensity ratio was calculated for each clone on the array and used as the normal level of expression for that particular clone. These values were compared to the BAC clone intensity ratios for each tumor sample to identify clones with intensities that were 3x standard deviations (s.d.) from the normal copy number. The clones with intensities less than or greater than this threshold were considered to be a loss or gain, respectively, of the genomic sequence represented by the clone. The 3x s.d. threshold was used as an initial first step to identify regions of copy number

**Figure 2.4**: This plots illustrates the copy number variation observed in a single CGH experiment of a grade III AO. The x-axis represents the position on the chromosome and the y-axis represents the $\log_2$ ratio of the clone intensity. The red horizontal line indicates a threshold of significance of clones considered to be losses and the green horizontal line represents clones considered to be gains. For this particular tumor sample, there are losses of chromosome 1p, the majority of 13, and 19 while there is a gain of chromosome 8p. Chromosome X showed a gain while chromosome Y was lost because of the hybridizing of DNA samples of the opposite sex. These chromosomes serve as an experimental control and confirm the sex of the DNA sample was correct.
variation and more complex methods were going to be used in subsequent analyses for comparison, but a change in research topics only allowed enough time for the analysis discussed in this work.

To identify intervals of chromosomal rearrangements, I identified genomic locations where: 1) there were at least 3 adjacent clones on the chromosome with log ratio intensity values at least 3 standard deviations from the normal log ratio, 2) the clones all had the same copy number variation (all gains or losses), 3) and the changes were observed in all tumor samples or samples from a particular subtype. These regions were defined as minimal critical regions (MCRs) to identify genes that were contributing to the overall OD phenotype and for each OD subtype. The identification of MCRs provided an approach that’s more targeted than the visualization described above.

2.4 Results

2.4.1 Summary of genomic variation for all 74 ODs

The most common copy number changes across all of the ODs were the loss of chromosomes 1p and 19q in 54 tumors, which is the signature genomic aberration associated with ODs. Genome plots were generated for each tumor sample and visually examined. The most commonly identified large scale losses were on chromosomes 4, 9, 13, 14, 15, and 18 while chromosomes 7, 8, 11, and 19 contained the majority of the gains observed (Figure 2.5). The overall progression of ODs can be associated with increased genomic variation among the tumors and the presence of multiple genomic changes in a single tumor. The genomic aberrations of each stage are described in more detail in the following sections.
2.4.2 Unclassified ODs

Of the 14 unclassified ODs, there were 8 tumors with the simultaneous chromosomal loss of 1p and 19q and of these, there were 3 tumors with no other substantial genomic aberrations. For the 5 with additional aberrations, there were only clusters of clonal deletions/duplications instead of large, contiguous blocks. The

\[\text{Figure 2.5: The vertical red and green lines on each chromosome indicate the number of tumors containing deletions or amplifications, respectively. The percentages indicate the number of individuals with a particular genomic aberration, e.g. 74% of the ODs have deletions of chromosome 1p. The boxes overlapping the chromosomes with a letter noted to the right of the box indicate minimal critical regions (MCRs) identified. The minimal regions were defined by a similar genomic aberration identified in at least 3 adjacent BAC clones. We identified these regions in ODs that had the chromosomal losses of 1p and 19q.}\]
locations of these deletions and duplications were as follows: deletion on 8p (-8p), -9p, duplications on 9q (+9q) and across 7, deletions across 10 (-10), -14q, -15, -18. One patient tumor sample contained 3 large-scale genomic aberrations, with a substantial amount of clonal gains on chromosome 7 and a complete loss of 10, which is a pattern associated with a more aggressive form of brain tumor, grade IV glioblastoma. The remaining 6 tumors did not have the chromosome 1p or 19q loss and only 2 of the patient samples had large-scale rearrangements. For one tumor, the majority of 15q and chromosome 18 were lost while there were clone duplications on 17q. The other tumor had losses of 6p, 14, and substantial clonal losses on 4q, 7p, and chromosome 16.

2.4.3 Grade II well-differentiated ODs (WDs)

There were 30 grade II well-differentiated ODs. Of the 30, there were 22 tumor samples that had chromosome losses of 1p and 19q and of these, there were 5 tumors with no other substantial genomic aberrations. The next most common aberrations among the tumors were wide-spread deletions on chromosomes 1, 4, 9p, 13 and 14 as well as gains on 11.

One OD only had the loss of 1p, but not the loss of 19q. It had very little copy number variation otherwise.

There were 7 tumors without the chromosomal loss of 1p and 19q. One tumor sample did not have any obvious aberrations while the remaining samples had a variety of aberrations. There was no obvious pattern to the genomic changes occurring in the tumor samples. Chromosome regions that had the most substantial change were deletions of 4q, 6p, 13p, and 14 with duplications of chromosome 5 and 8. The majority of the changes in the WDs were deletions, with a substantial amount of clonal gains on 7, 9p and 10p.
2.4.4 Grade III anaplastic oligodendroglioma (AOs)

A total of 30 AOs were analyzed using aCGH. Of the 30 patient samples, 24 had chromosomal losses of 1p and 19q. There was one tumor that had the 19q deletion without the loss of 1p that had several other chromosomal deletions including 5q, 10q, 13, and 21p, in addition to duplications of 8q and 12p. The most common large-scale genomic changes were deletions of 4, 9p, 15, and 13. Numerous studies have observed these aberrations among ODs where the loss of 9p is associated with AOs [109] as well as tumor necrosis and microvascular proliferation that are two mechanisms commonly observed in the more malignant types of brain tumors. The majority of the tumor samples had at least 2 chromosomal alterations, but there did not appear to be any particular pattern of deletions or amplifications.

2.4.5 Evaluation of clone-based genetic variation

To provide a quantitative measure of the genomic changes across the different tumor groups, I counted the number of individual clones demonstrating either a deletion or amplification. Groups were defined by tumor grade (unclassified, grade II, grade III) and the loss of 1p/19q (Table 2.1). Table 2.1a shows the number of genomic aberration for each tumor group. In the case of the unclassified tumors, which were assumed to include grade II and grade III tumors, 72% of all aberrations represented deletions. For the grade II tumors, 68% of the aberrations were deletions while only 43% were deletions in the grade III tumors. The unclassified tumors had the lowest number of genomic changes of all tumor groups. There was a significant increase in genomic aberrations in grade III tumors as compared to grade II tumors (Chi squared, p < .0001). It appears that there is a more substantial number of deletions in the transition from grade II to grade III tumors.

When the deletions and duplications in the tumor groups with and without the
chromosomal loss of 1p/19q were evaluated (Table 2.1b), there were fewer genomic changes observed in the tumor groups without the 1p/19q loss than those with the deletion for all tumor stages, but both tumor groups had increasing numbers of genomic aberrations with tumor progression.

Table 2.1: A) The total number of clone deletions, duplications, and total number of aberrations observed for each tumor group is listed. The unclassified ODs had the least number of changes while grade III ODs had the most. The number of duplications had the largest increase from grade II to III when compared to the number of deletions. b) The total number of clone deletions and duplications in ODs with and without the chromosomal loss of 1p and 19q (1p/19q) for each tumor type. For all tumor types, there was a larger number of both types of aberrations in the ODs with the 1p/19q loss as compared to those ODs without the loss as well as from grade II to III ODs. Again, the number of aberrations observed for both 1p/19q groups were less than the grade II and III ODs.

2.4.6 The identification of minimal critical regions (MCRs)

The tumors that did not have the chromosomal loss of 1p and 19q were of particular interest because they did not have the characteristic alterations typically
associated with the ODs yet they were still classified as such, so there may be other genes involved in the formation of these tumors that influence their OD phenotype. To determine the chromosomal changes contributing to the formation of the tumor, MCRs were identified. The MCRs consisted of chromosome changes (deletions/duplications) in at least 3 adjacent clones showing a substantial copy number change across all of the tumors under consideration. There were 20 ODs lacking the 1p/19q deletions (6 unclassified, 8 grade II, and 6 grade III). The investigation of 14 MCRs identified in these 20 ODs resulted in the discovery of several genes associated with the cell growth cycle and tumor formation. Genes identified in these regions associated with cancer pathways were the Rab34, FAT, SHH, MDM1, MDM2, CDKN1A, and Raptor. PDGFRA was also identified which has been implicated in GBM tumorigenesis.

2.4.7 Discussion

We identified genomic aberrations associated with the overall OD phenotype and the tumor subtypes. These changes were then correlated with tumor stage and progression. The characteristic chromosome loss of 1p and 19q associated with ODs was the most common genomic change identified across all 74 tumors analyzed. This was also the most common genomic alterations in the grade II tumors which provides additional evidence that this alteration is an early event in tumor formation, as demonstrated in [110][111]. The loss of chromosome 1p and 19q has also been associated with increased tumor sensitivity and better response to chemotherapy [96]. Several studies have identified tumor suppressors, such as CDKN2C and p53, located on chromosome 1p and the loss of these genes can have a direct impact on tumor formation [112][113][114][115]. Additionally, this may also be a result of a cooperative effect of the two chromosome losses that may disrupt the tumors aggressive potential or gene functions that may reduce the tumor cells’ viability. This could be
due to tumors progressing from grade II to III acquiring new deletions or that cells with deletions in grade II die before the tumor progresses to stage III.

There were increases in the total number of alterations from grade II to III. In particular, there were more deletions than amplifications in grade II while there were more amplifications in grade III. The initial stage of tumor formation appears to be associated with the loss or malfunction of genes while the later more severe stage appears to be associated with increased copy number. The presence of amplifications in AOs has been identified as a poor indicator of survival [116] which is intuitive with the diagnosis of more aggressive tumors.

The MCRs identified in the tumors with the 1p19q loss provided evidence of genes that may be implicated in the tumorigenesis of oligodendroglioma. The most noted genes in the group of potential candidate genes are TP53 and MDM2. The inhibitory and stimulatory relationship between MDM2, an oncprotein, and TP53, a tumor suppressor, has been shown to promote cell homeostasis, but the disruption of either permits unregulated cell growth and inevitably leads to cancer [117][118]. Interestingly, MDM2 and TP53 interact with the p53 gene as a binding protein and inhibitor. In the case of MDM2 being upregulated, the expression of p53, which is a tumor suppressor, would enhance cell proliferation in the tumors and aid in the development of ODs. The deletion of TP53 in the tumors, in concert with MDM2, would promote uninhibited activity of this tumor that could provide a substantial effect on tumor growth. The overexpression of MDM2 and TP53 mutation have been studied in the more aggressive tumors, glioblastomas, and a characteristic trademark of these tumors [119][85]. This could suggest that the mechanism involved with these genes could be strong enough to drive the development of these tumors given that they do not have the characteristic 1p/19q loss typically associated with these tumors. Alteration of the PDGFRA gene has been implicated in the pathogenesis of glioblastoma, the most malignant type of brain tumor [120]. RASA2 is a GTPase
activator of p21 that was found to be duplicated and deleted in 4 tumor samples. The deletion of the RASA2 gene could cause increased cell proliferation due to the inhibitory effect of p21 on the cell cycle, which is important in the prevention of cancer formation [121]. The amplification of the MYCN and FEV genes, which were previously identified as oncogenes [122][123], could influence the development of tumors while GATA6, FOXM1, SHH, WDR37, and RAD52 are involved in cell cycle maintenance tasks, such as apoptosis and cell proliferation, that could enable the uncontrolled growth of cells and lead to tumorigenesis.

The ODs analyzed in this study were classified using histological analysis, but some of the tumors could not be classified by this method alone, as illustrated by the group of tumors denoted as "unclassified ODs". The heterogeneous nature of ODs often makes it difficult to classify these tumors and differentiate between tumors derived from the same lineage, such as oligo-astrocytomas, that are known to have the 1p19q loss and have similar histological profiles [124]. The presence of the 1p19q loss in other tumors could also influence the misclassification of tumors as ODs. It has been shown that ODs without the 1p19q loss still retain the usual histological appearance associated with ODs [109]. For these reasons, evaluating the genomic profiles of ODs for further classification is important given that the presence of additional aberrations could identify more aggressive tumors that is critical to the proper diagnosis and treatment of patients. An example of this was found among our ODs where one tumor sample did not have the characteristic 1p19q loss, but had a nearly complete gain of chromosome 7 and complete loss of chromosome 10, which are genomic aberrations associated with a more aggressive and deadly brain tumor, grade IV glioblastoma. Both amplifications and loss of chromosome 10 have been associated with poor prognosis in patients; therefore, it is important for the genomic properties of ODs to be known in the earlier characterization of the tumors.
to provide proper patient diagnosis and treatment.

Although CGH is useful for the identification of large-scale CNV in ODs, it cannot provide quantitative measures of the level of copy number changes observed in the tumors. Therefore, it is important to conduct follow-up studies to confirm the level of CNV using other experimental techniques, such as microsatellite analysis or fluorescence in situ hybridization (FISH), which will provide a more targeted approach to evaluating specific genomic regions. This will also ensure that there is truly a loss or gain of a genomic region and that it is not an artifact of poor hybridization or array quality using methods such as microsatellite analysis. Additionally, conducting studies to evaluate the effect of copy number variation on gene expression, using techniques such as gene expression or RT-PCR, will be important to determine if the functions of the identified genes are truly affected. The fact that epigenetic mechanisms also play a role in the development of cancer is another aspect of CNV in ODs that cannot be taken into consideration using CGH. This is another important mechanism that should be taken into account in these studies because instead of a copy number change in a genomic location. With all of these technical and biological issues taken into consideration, CGH can be a useful approach to the further classification and diagnosis of ODs, but it if more advantageous to use this experimental platform in conjunction with other methods to obtain the most accurate results.
Conclusions

The goal of my thesis work was to study the association between genomic variation and disease using two case studies: single nucleotide polymorphisms and their association with susceptibility to *Staphylococcus aureus* (*S. aureus*) infection in mice and copy-number variation and its association with tumor progression in the context of ODs. The characterization of SNPs in *S. aureus* infection was conducted using the ISHM while CGH was used to characterize the CNV in ODs. The studies described here utilized methods that provided additional insight into the pathogenesis of both diseases and could potentially have long-term clinical and methodologic implications for patient diagnosis and the development of novel therapeutics.

In the first study, we conducted in silico haplotype mapping using a modified version of the Pletcher et al. [52] method, with the primary modification being the approach for estimating Type I error. We conducted genome-wide permutations and estimated genome-wide Type I error from the distribution of the largest test statistics observed genome-wide. The advantages of this approach include that it accounts for the genome-wide nature of the test for association and the lack of independence between blocks (and test statistics) used ISHM to identify genomic regions associ-
ated with susceptibility to *S.aureus* infection in mice. When we previously applied the standard ISHM method in [52] for the kidney phenotype, we obtained 38,172 genomic regions significantly associated with the bacterial load in the kidney and 2,742 genes were identified among this set of regions. This suggests that we have improved upon the Type I error rate observed in the initial method, but our approach also has the potential to be too stringent in its analysis given that we only obtained a single haplotype block.

In regards to the localization of the method, it proved to be fairly accurate, but this is only speculation because we did not rigorously assess the effect of the common strain ancestry on the method. Therefore, this approach could be biased toward the genetic composition of the mice selected where the haplotype identified is an artifact of this genomic property. We could address this issue in future studies by conducting additional alternative simulations and selecting additional causal SNPs that have different allele frequencies to test how well the method localizes.

Despite the limitations of our modified ISHM approach, it has identified potential susceptibility genes in the context of *S.aureus* using genetic variation in inbred strains of mice that appear to be biologically relevant to *S.aureus* infection. We identified a single block on chromosome 7 associated with colonization of the kidney by *S.aureus* in infected mice. Using a simulation study, we extended this region by 380kb resulting in a region that encompassed 40 genes, 26 of which are kallikrein genes. Using gene expression data, we prioritized the genes contained in the significant region by identifying genes that were differentially expressed between susceptible and resistant mouse strains that resulted in 8 candidate genes. The gene expression data provided biological validation for the genes identified by ISHM and provided support for their role in host susceptibility to infection. Additionally, with the role of the KKS system as one of the first lines of defense in infectious disease, we can speculate that the kallikrein genes are contributing to susceptibility to *S.aureus* infection.
The functions of the kallikrein genes in the context of *S. aureus* infection will need to be experimentally validated, in the mouse, to determine whether the genes are truly associated with susceptibility to *S. aureus* infection before drawing any final conclusions. Following validation of the genes’ functional roles, these genes will have to be evaluated in humans, more specifically *S. aureus* infected patients, to determine if the predicted association of the kallikrein genes with *S. aureus* infection is real. Nonetheless, the findings of this study appear to be promising based on the importance of the KKS in the innate immune response to bacterial infections and lines of evidence supporting its role in susceptibility to disease.

For the characterization of ODs, I conducted array-based CGH on 74 oligodendroglioma (OD) to identify patterns of genomic variation associated with this tumor type and its two subtypes: grade II ODs and grade III ODs. Additionally, CGH was conducted on 27 normal, or cancer free, DNA samples to determine the normal copy variation in disease free individuals, which was then used to identify smaller genomic intervals with substantial copy number differences, known as minimal critical regions (MCRs), when compared to the OD samples.

Using array based CGH to evaluate the genomic variation associated with ODs proved to be useful in the characterization of the brain tumor phenotypes. When all 74 ODs were analyzed, we found that 54(73%) of the tumors had the characteristic chromosome loss of 1p and 19q, which provides additional evidence that these tumors are ODs and support for this loss being an early event in tumor formation. ODs with this pair of aberrations are associated with a better prognosis and response to treatment, which could not be assessed in this particular study, but in future studies this clinical information could be added to this analysis to see if this hold true for these tumors.

The group of unclassified ODs is an example of the primary limitation of histological analysis, which relies on visual inspection of cell morphology. The heterozygosity
of ODs or a tumor composed of more than one tumor type can easily be misclassified if some of the tumor cells are not detectable by eye [86]. When evaluating the genomic profiles of these tumors, they were distinctly different, which indicates the advantage of using this approach to characterize the tumors. The unclassified tumors were considered to be ODs where the grade was unknown, but some of the tumors had aberrations similar to the more aggressive tumor types that were either a mixed type of OD or a different tumor altogether. Therefore, using an approach like this in addition to histological analysis could reduce the misclassification rate of these tumors. Future studies could test the utility of using this method by classifying a group of tumors by histological analysis, with and without genomic profiling of the tumors, and assess the accuracy of the predictions. This could also be expanded to analyzing different types of tumors to test the ability of these methods to differentiate between the type of tumors.

Characterizing ODs by their genomic profiles could also be useful in patient diagnosis and a more specialize treatments by correlating the observed chromosomal changes with each patient for the development of more specialized treatment. A similar strategy was used to develop specialized treatments for patients with advanced stage ovarian cancer by using gene expression profiles to identify patients that were most likely resistant to particular forms of treatment [125]. This type of strategy could also be useful in the identifying more effective treatments used to treat ODs based on the alterations present in the tumor, especially in the case where the tumor is histologically diagnosed as low grade, but according to its genomic profile it contains alterations that signify a more aggressive type of tumor.

The identification of MCRs in the ODs provided a list of interesting genes for each tumor group. For those tumors with the 1p19q loss, several genes previously identified in cancer studies were discovered, which illustrates that CGH is useful for identifying genes associated with CNV that could influence tumor formation and
progression. The drawback of this method is that large BAC clones that contain several genes will make it difficult to identify which gene is affected if only a portion of the region has a substantial copy number change. The more recent platforms designed for CNV analysis, such as SNP and oligo-nucleotide arrays, improved upon this issue by using smaller hybridization targets which may be more appropriate for studies that require the detection of small regions of CNV, otherwise CGH is well suited for the identification of large-scale aberrations.

Future analyses for the CGH analysis could be extended to include clinical information about the tumors to utilize patient information to determine the influence of the identified genomic aberrations on clinical outcome. From a computational standpoint, the detection of regions of copy number variation could also be improved given that we used an arbitrary threshold to determine copy number changes. With the natural genetic variation across individuals, it would be necessary to account for this due to varying levels of CNV across the genome.

Both of the studies discussed in this work have demonstrated that genetic variation plays important roles in host susceptibility to \textit{S. aureus} infections and tumor progression and formation in ODs. They also illustrate the difficulties in localizing genetic variants responsible for the phenotype. The association of genetic variation with disease phenotypes can be a valuable tool for a better understanding of disease pathogenesis and the identification of genetic factors that are the most influential in particular disease states. In the case of SNPs in \textit{S. aureus} infections, individuals could be screened for a particular allele associated with the susceptible group and provided with preventative treatments to minimize any complications if they become infected. In the case of ODs, the availability of the genomic profiles of tumors would enable clinicians to make well-informed decisions and aid in improved patient diagnoses and individualized treatments. Therefore, the study of genetic variation in disease not only provides a better understanding about the genetic factors involved
in pathogenesis, but it also the chance to improve the quality of health care.

Despite the utility of the methods employed, both computational methods have limitations. In the first study of *S.aureus* infection, we only detected a significantly associated haplotype block for one of four phenotypes. This is likely due, at least in part, to the fact that the ISHM approach assumes that phenotypic similarities between moust strains result from underlying genetic variants common to the set of phenotypically similar strains. We did not rigorously account for strain relatedness outside of the global $w_g$ included in the F-statistic. Since the choice of inbred strains used in a study can influence the results obtained from the analysis, we may have limited our ability to detect associations due to limited genetic diversity across this particular set of strains. Therefore, in the case that more strains are analyzed, we may be able to detect more significant associations across all four phenotypes. Additionally, our study phenotypes (e.g., survival, serum IL-6) may result from interactions between many different genetic variants, each of which is shared by only a subset of the strains exhibiting similar phenotypes. Second, we limited our analysis to the inbred strains for which resequenced data are available from the Perlegen database. While the use of imputed data could increase the number of strains used, which is important for phenotypes with a small effect size and for which there is low power, the use of imputed data could also increase the likelihood of spurious associations, in particular when imputation error is correlated with strain relatedness. We therefore opted in this study to focus on strains for which resequenced genotype data are available. It is likely, however, that with a larger mouse strain panel, additional associations would be detected. Finally, the survival times are right-censored, in that the true median survival time for the resistant strains falls outside of the monitoring period. Thus, the median survival times for these strains are underestimated. This results in an underestimate in the F-statistic sum of squares for the within-and between-haplotype group variability. While underestimates of the within-group
variability may result in false positive associations, underestimates of the between group variability may result in false negative associations.

In the study to characterize various grades of OD brain tumors, the approach used to detect genomic aberrations was limited by the method used to detect genetic aberrations. The median fluorescent intensity ratio calculated across the 27 normal CGH experiments could easily be influenced by the natural genetic variation observed in the normal sample with the median that could result in a false positive, or negative when compared to the disease samples. For instance, in the region where there is a CNV in the normal sample with the median, there could appear to be a copy number change in the disease sample or a copy number change in the tumor sample could be eliminated because there would appear to be no copy number change between the two. Therefore, it would be necessary to calculate a value that captures the natural genetic variability across the samples for a more appropriate basis of comparison. Additionally, there was no statistical model in place to appropriately compare the two data sets or correct for batch effect, which is the experimental variation that could occur, that could also result in false positives.

In spite of these important limitations, we were able to identify a genomic region significantly associated with susceptibility to \textit{S. aureus} infection, for which there is strong supporting experimental evidence, and which contains a large gene family whose members are promising candidate genes for future biological validation. Future studies in the mouse will identify specific members of the family with a role in \textit{S. aureus} pathogenesis in the murine host and will elucidate the specific mechanisms by which the gene products confer susceptibility. These studies in the mouse will be followed by studies to evaluate the role of the corresponding orthologous genes in human susceptibility to \textit{S. aureus}. We were also able to demonstrate the utility of using CGH to generate the genomic profiles of ODs and the importance of using this information in the classification and diagnosis of patients. Although these stud-
ies are only initial steps in fully understanding the underlying genetic mechanisms influencing disease pathogenesis, the further development of these types of methods would provide significant contributions to this endeavor.
Bibliography


77


[97] C. Ramirez, C. Bowman, C.-A. Maurage, F. Dubois, S. Blond, N. Porchet, and F. Escande, “Loss of 1p, 19q, and 10q heterozygosity prospectively pre-


Chromosome 1 and Rare Mutation of the CDKN2C Gene from 1 p32 in Oligodendroglial Tumors,” *Journal of Neuropathology & Experimental Neurology*, vol. 58, no. 10, p. 1041, 1999.


Biography

Nicole Verteree Johnson was born on August 10, 1981 in Waynesburg, Pennsylvania. She received her Bachelor of Science (B.S.) degree from West Virginia University in 2004. She then began her graduate training at Duke University in 2004 as a graduate student in the Computational Biology and Bioinformatics Program and completed her Doctorate of Philosophy (Ph.D) in August 2010. After graduation, she will be working at the Novartis Institute of Biomedical Research in Cambridge, MA as a Presidential Postdoctoral Fellow.